The role of prosody and voice quality in text-dependent categories of storytelling across languages

Raúl Montaño, Francesc Alías

GTM - Grup de recerca en Tecnologies Mèdia, La Salle - Universitat Ramon Llull
Quatre Camins, 30, Barcelona, Spain
raulma@salleurl.edu, falias@salleurl.edu

Abstract

In contrast to full-blown emotions, storytelling speech entails a particular speaking style that contains subtle expressive nuances of which little is known. In the present work, we study the role of prosody and voice quality while searching for cross-linguistic acoustic similarities in two categories of storytelling speech that are defined by their lexical components: the descriptive mode and sentences that specify a character intervention, together with a third neutral category (perceptually validated as reference). The study addresses four narrators using four different European languages (English, French, German and Spanish) expressing the same story. After conducting several statistical and discriminant analyses, we find that all narrators under analysis exploit some acoustic parameters in a similar way to differentiate among the analysed storytelling categories. Specifically, we observe that three prosodic features (mean fundamental frequency, mean intensity and number of silent pauses) and two voice quality parameters (mean Harmonic-to-Noise Ratio and Maxima Dispersion Quotient) explain a relatively similar proportion of the variance among storytelling categories in all languages. Moreover, the classification results obtained from the discriminant analysis are comparable for the three considered storytelling categories across languages.

Index Terms: storytelling, speech analysis, prosody, voice quality, cross-language

1. Introduction

Among myriad speech analysis studies devoted to expressive speech, the majority have been focused on the analysis of emotions (see [1] and references therein). In contrast, the storytelling speaking style, which is an expressive style per se (specially, when aimed at children), has not received so much attention in the literature. In addition, studies focused on storytelling speech have followed different analysis approaches, ranging from very specific expressive aspects of storytelling speech like suspense situations [2] to more general studies using an annotation based on structure of tales [3, 4]. Moreover, basic emotions have also been linked to storytelling revealing some contradictory results when compared to previously reported emotional acoustic profiles in the literature [5]. Probably, relating the storytelling indirect discourse (see section 2 for more details) to basic emotions is inappropriate because the narrator is not self-experiencing the emotions, but trying to entertain the audience and engage them in the story [6]. Therefore, the indirect storytelling speech seems to be characterized by more subtle expressive nuances compared to typical emotions. In this sense, while prosody has proven crucial in storytelling speech [2–4], it has only been suggested that voice quality (henceforth VoQ) may also play an important role in this speaking style [2, 3]. Certainly, in the same way VoQ is relevant for emotions [7], it may be of great interest for the characterization of subtle affective variations like those present in storytelling speech. Furthermore, up to our knowledge, there are no cross-language studies focused on acoustic analysis of storytelling speech categories, although certain emotions and speaking styles have been analysed in a cross-language context. For instance, emotions have displayed many similarities in their acoustic-perceptual properties across different languages [8, 9], whereas certain speaking styles (e.g., polite and informal speech) have also shown several acoustic "tendencies" among speakers of different languages [10].

In this paper, the indirect storytelling speech is acoustically analysed with the objective of exploring the role of prosody and VoQ in some storytelling categories across four European languages: English, French, German and Spanish. Concretely, we analyse two storytelling categories that are defined by their lexical components: the descriptive mode and sentences that specify a character intervention (denoted hereafter as post-character sentences). In addition, a neutral category is also analysed as reference after being perceptually validated. We also check if each of these text-dependent categories show particular acoustic cues in a similar way across languages.

This paper is structured as follows. Section 2 explains the storytelling categories addressed in this work. Next, Section 3 introduces the prosodic and VoQ parameters taken into account for the subsequent analyses. Then, Section 4 explains all the experiments together with the obtained results. Finally, this paper ends with conclusions and future work in Section 5.

2. Categories in storytelling speech

Storytelling speech has been analysed in the literature from diverse points of view. Some works have followed an annotation based only on the identification of structural and lexical elements of text [3, 4]. Nevertheless, other works have evidenced that perceptual analyses may also be necessary to identify categories containing particular expressive nuances [2], which are not directly inferred from the text-level analysis.

Tales and stories typically contain narrative, descriptive and dialogue modes [11]. The narrative mode is generally used to inform the listener/reader about the actions that are taking place in the story, whereas the descriptive mode has the function of describing characters, environments, objects, etc. On the other hand, the dialogue mode is present when the characters have a conversation and their turns explicitly appear in the story. Therefore, the narrative and descriptive modes belong to the indirect discourse, whereas the dialogue mode represents a direct discourse. As a preliminary step to look for cross-language

similarities and/or differences in terms of prosodic and VoQ patterns, we focus on three storytelling categories based on storytelling discourse modes [4].

Particularly, our study deals with categories showing specific lexical cues that can be annotated from text [4, 11]. In what concerns text-dependent storytelling categories, at least two are widely used [4, 12]: descriptive and post-character sentences. On one hand, an important lexical characteristic of the descriptive mode in stories and tales is the large number of adjectives used. Moreover, verbs like "to be" and "to have" in the past and present tenses abound along this mode (e.g.,*"He was a tall, strong boy with faded bluish eyes"*). On the other hand, usually right after characters' interventions, the narrative mode contains post-character sentences that specify the character turn. These sentences show specific lexical cues, e.g., they usually start with a declarative verb in the third person ("said", "answered", "murmured", "asked", etc.) [12].

Finally, in line with most affective speech analysis and synthesis studies [1], we also consider a reference neutral category. Within the narrative mode of tales and stories, there are merely informative sentences about actions or facts containing neutral lexical elements (e.g., *"The boy came into the living room"*). While these sentences could be identified via text only, a posterior perceptual validation step is included to ensure their neutral expressiveness (see Section 4.1). This way, those sentences with non-neutral expressiveness (e.g., since the storyteller has added a suspenseful style [2]) are discarded to avoid biasing the comparisons with the other two categories.

3. Prosodic and VoQ parameters

The following acoustic features describing prosodic and VoQ information are considered for the acoustic analyses of the speech corpora at hand, as they have proven useful in previous works devoted to affective speech analysis [13–15]:

- **Fundamental frequency**: f0 mean ($f0_{mean}$) and f0 interquartile range ($f0_{IQR}$) in Hz.
- **Intensity**: Mean intensity (int_{mean}) in dB.
- **Tempo**: Speaking rate (SR) in syllables per second (excluding pauses) and number of silent pauses (Nsp).
- **Jitter**: Cycle-to-cycle variations of the fundamental period. Jitter local from Praat in % [16].
- **Shimmer**: Cycle-to-cycle variations of the waveform amplitude. Shimmer local from Praat in % [16].
- **Harmonic-to-Noise Ratio (HNR_{mean})**: Relation between the energy of the harmonic part and the energy of the rest of the signal in dB [17].
- **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 [18].
- **Hammarberg Index (HammI)**: Difference between the maximum energy in the band frequencies [0, 2000] Hz and [2000, 5000] Hz expressed in dB [19].
- **H1H2**: Difference of amplitude between the first two harmonics in dB [20].
- **Spectral Slope (SS)**: Spectral slope computed with the energy band difference function of Praat [16] (in dB).
- **Normalized Amplitude Quotient (NAQ)**: It describes the glottal closing phase using amplitude-domain measurements [21].
- **Parabolic Spectral Parameter (PSP)**: It describes the spectral decay of the glottal flow with respect to the maximal spectral decay [22].
- **Maxima Dispersion Quotient (MDQ)**: This parameter measures how impulse-like the glottal excitation is via wavelet analysis of the linear prediction residual [14].

All these parameters were extracted using a Praat script specifically developed for this task [16], except for the glottal flow parameters (NAQ, PSP and MDQ), which were extracted using COVAREP (version 1.3.1) algorithms [23]. We kept the results only from vowels, as they represent stable voiced zones of running speech from which reliable acoustic parameters can be extracted [7, 13, 24]. The segmentation of the speech corpora into words, syllables and phonemes was carried out with the EasyAlign tool [25] for Spanish and French. For the German corpus we used the WebMAUS service [26], whereas for the English corpus we used the SPPAS tool [27]. All these automatic segmentations were manually corrected (if necessary) afterwards in order to dispose of reliable data (nearly 64,000 phonemes were revised).

4. Experiments

4.1. Speech corpora selection

In the experiments, we used four audiobooks where the same story in four different languages (English, French, German and Spanish) is interpreted by four native professional male storytellers. The story belongs to the fantasy and adventures genres, with children and pre-teenagers as its main target audience. Each audiobook contains approximately 20 minutes of indirect storytelling speech, composed of 250 sentence-level utterances. The text-level annotation with no audio input of the Spanish version of the story was entrusted to two expert annotators (with Spanish as native language) after a thorough briefing of the annotation goal. Sentences where the annotators did not agree were discarded for the analyses. The inter-rater agreement was found to be Kappa $\kappa = 0.704 (p < 0.001)$, due to some disagreements in several sentences within the neutral sentences annotation (no disagreement in descriptive and post-character sentences). From the resulting 250 sentences of the Spanish version, 24 were labelled as descriptive sentences while 40 belonged to the post-character category, which can be also borrowed from the rest of languages as we work with parallel corpora. Regarding neutral category, 55 sentences were extracted initially from text. Next, as introduced in Section 2, a two-stage perceptual validation study was also carried out in order to ensure the neutral expressiveness. Firstly, the two experts perceptually validated the uttered neutral sentences of the Spanish version. As a result, the number of valid neutral exemplars was reduced to 45 due to some disagreements between the two experts ($\kappa = 0.614, p < 0.001$). Then, the neutrality of the same sentences uttered in the remaining languages was validated taking the Spanish corpus annotation as reference to obtain a representative parallel neutral corpora, besides having a first glance at cross-language similarities.

To that effect, 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 [28]. In this case, we recruited 18 native Spanish speakers (14 males, 4 females; mean age: 33 ± 8.6) to take the tests in order to reinforce the cross-language analysis. In these tests, we took into consideration the fact that, although listeners perform best when listening to speakers of their native language, they perform well at perceptually identifying neutral utterances when produced by speakers of a foreign language [29].

The corresponding Spanish neutral audio reference was presented in each step together with the utterance to be evaluated

Table 1: Wilks' lambda values for each parameter per language (ENG: English, FRE: French, GER: German, SPA: Spanish). The asterisk (*) indicates p<0.05 in the ANOVA among categories while a double asterisk (**) indicates p<0.01.

Parameter	Wilks' Lambda			
	ENG	FRE	GER	SPA
SR	0.814**	0.810**	0.674**	0.938
Nsp	0.539**	0.716**	0.693**	0.714**
$f0_{mean}$	0.516**	0.912*	0.660**	0.514**
$f0_{IQR}$	0.815**	0.928*	0.925*	0.681**
int_{mean}	0.727**	0.998	0.813**	0.776**
Jitter	0.855**	0.983	0.756**	0.924*
Shimmer	0.919*	0.974	0.895**	0.931*
HNR_{mean}	0.605**	0.957	0.662**	0.594**
pe1000	0.994	0.988	0.916*	0.856**
HammI	0.942	0.928*	0.962	0.918*
SS	0.928*	0.952	0.901*	0.789**
NAQ	0.944	0.918*	0.784**	0.912*
PSP	0.916*	0.996	0.756**	0.833**
MDQ	0.686**	0.786**	0.811**	0.821**
H1H2	0.641**	0.833**	0.894**	0.976

(the same sentence uttered in other language). The evaluators could listen to both audio signals as many times as they wanted and they had to answer the question *"The expressiveness of the audio under evaluation compared to the expressiveness of the reference audio is:"*, choosing among three possible answers: *"Higher"*, *"Roughly the same"*, *"Lower"*. We randomly selected 30 Spanish neutral utterances from the original corpus of 45 sentences, in order to avoid user fatigue while maintaining balanced speech corpora for the 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, in this case we used the free-marginal Kappa as inter-rater agreement measure [30]. This method derived from the Fleiss' fixed-marginal multirater Kappa [31] avoids the prevalence and bias paradoxes of the fixed-marginal solution [32]. The obtained free-marginal Kappa values were $\kappa_{free} = 0.78$ for the English test, $\kappa_{free} = 0.80$ for the French test and $\kappa_{free} = 0.81$ for the German test. At this level of κ_{free}, the agreement is usually deemed as "substantial" [33]. Finally, an exemplar was defined as an utterance with proportion of agreement per item [31] greater than 0.61, showing substantial agreement on choosing the option *"Roughly the same"*. As a result, 27 neutral utterances for each language were considered for the subsequent analyses.

It is worth noting that the averaged value across tests of proportion of category assignment [31] resulted in 0.92 for the *"Roughly the same"* category. Thus, this result together with the substantial values of κ_{free} are a first encounter of cross-language similarities in storytelling speech, since narrators used a similar expressiveness for the majority of neutral sentences under perceptual evaluation. In Section 4.3, objective acoustic similarities across languages within storytelling categories are investigated through statistical and discriminant analyses.

4.2. Cross-language acoustic analysis methodology

The cross-language analysis methodology is the following. Firstly, relative acoustic differences among post-character, descriptive and neutral storytelling categories are studied within each language. Then, we evaluate to what extent these relative patterns are present across languages, avoiding this way speaker-dependent and language-dependent acoustic profiles [34]. Similarly to previous cross-language studies [8, 9], we conduct a series of statistical and discriminant analyses (using the statistical software SPSS [35]) in order to assess if the different storytelling categories under analysis can be acoustically differentiated. As a first step, we perform a multivariate analysis of variance (MANOVA) within each language (using Pillai's Trace test statistic), considering all acoustic parameters as dependent variables. Then, a series of univariate analyses are conducted to evaluate differences among categories for each parameter. To conclude the statistical analyses, Tukey's Honestly Significant Difference (HSD) post-hoc tests results (i.e., pairwise comparisons between categories) on parameters showing the largest cross-language similarities are discussed. Finally, a discriminant analysis is also carried out within each language to assess how the different storytelling categories can be discriminated based on the acoustic parameters taken into account. Wilks' lambda is reported as a measure of discriminating capability of each parameter, as the smaller this value the more important the parameter to the discriminant function.

4.3. Results and discussion

After conducting the MANOVA, we found statistically significant results for each language: English ($F(30, 150) = 4.634, p < 0.001$), French ($F(30, 150) = 3.752, p < 0.001$), German ($F(30, 150) = 3.618, p < 0.001$) and Spanish ($F(30, 150) = 3.609, p < 0.001$). These results confirm the existence of one or more significant mean differences among storytelling categories (when considering all acoustic data) for each language. In order to refine this finding, we conducted a series of univariate analyses. One-way ANOVA's significance results for each parameter are represented in Table 1 together with Wilks' Lambda values from the discriminant analyses. As it can be observed in Table 1, Nsp, $f0_{mean}$, $f0_{IQR}$ and MDQ show statistically significant differences in one or more pairwise comparisons between categories in each language. Moreover, the English, German and Spanish narrators also exhibit similar results in int_{mean}, jitter, shimmer, HNR_{mean}, SS and PSP. Nevertheless, these parameters do not reach any statistically significant difference among categories in the French narrator. This fact implies that he uses similar expressiveness for the three storytelling categories, showing the least expressive variability among the gathered narrators.

Before describing post-hoc comparisons results, we investigated which parameters showed interesting cross-language similarities in discriminating among categories [8, 9]. For each narrator the discriminant analysis showed a major canonical function accounting for the majority (75-92%) of variance across categories. This function was in every case strongly correlated with $f0_{mean}$, while in three major canonical functions Nsp, int_{mean}, HNR_{mean} and MDQ manifested as good predictors too. For that reason, we consider that these parameters are the ones showing the largest cross-language similarities in terms of intra-language relative differences, besides showing a good cross-category discrimination capability (see Wilk's lambdas in Table 1). Therefore, further inspection of Tukey's HSD post-hoc tests was conducted on these five key parameters to assess their behaviour among the considered storytelling categories and languages.

For all aforementioned key parameters, we obtained larger values in the descriptive utterances and lower values in post-character utterances when compared to their neutral category

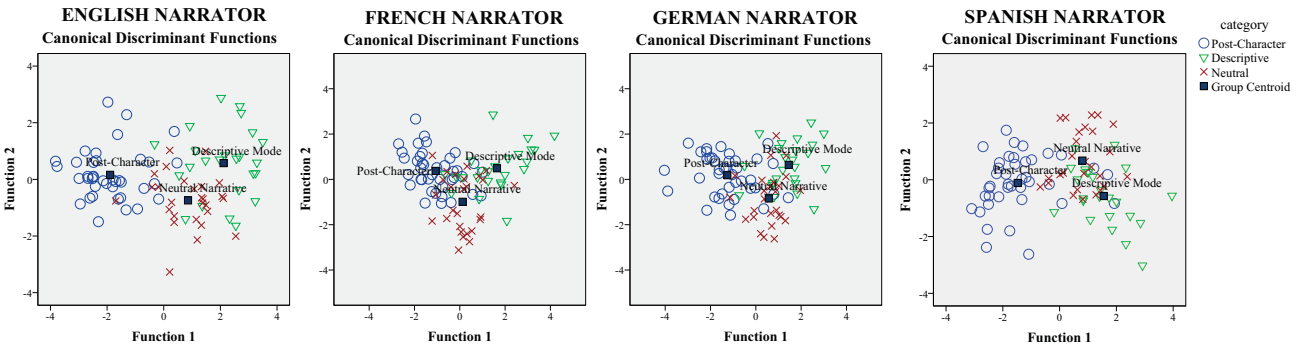

Figure 1: Combined-groups plots of the discriminant analysis for each language.

Table 2: *Cross-validated grouped cases correctly classified per language: English (ENG), French (FRE), German (GER) and Spanish (SPA) (in %).*

Category	ENG	FRE	GER	SPA
Post-Character	82.1	72.5	72.5	77.5
Descriptive	70.8	58.3	58.3	54.2
Neutral	59.3	55.6	66.7	51.9
Averaged	72.2	63.7	67.0	63.7

counterparts. Nonetheless, Tukey's HSD post-hoc tests do not mark all these differences as significant. Regarding $f0_{mean}$, post-character utterances show a significantly lower $f0_{mean}$ value than neutral and descriptive utterances in all languages except for French, whereas differences between the descriptive and neutral categories are not significant. In what concerns int_{mean}, it follows a similar pattern among all cases with the exception of the French language. Specifically, post-character utterances are in all cases significantly lower in intensity with respect to the neutral reference, while descriptive sentences are uttered with a similar intensity to such reference. Concerning Nsp, descriptive utterances show a significantly larger number of pauses than the rest of categories for all languages, although differences between post-character and neutral utterances are not significant. It seems logical that narrators use a slower tempo in descriptive utterances, as the information present in these sentences is rich and important. However, this slower tempo with respect to the rest of categories is only maintained for all languages in Nsp in contrast to SR. On the other hand, the MDQ is significantly lower in post-character utterances than the other two categories for all narrators except for the French narrator, which again does not show a significant difference when compared to the neutral category. However, he is the only narrator with significantly higher MDQ value in descriptive utterances with respect to the neutral reference. These MDQ results show a tendency of narrators to use a breathier voice in descriptive utterances and a tenser voice in post-character utterances [14].

From the acoustic analyses, two main general observations can be drawn. Firstly, English, German and Spanish narrators show several significant cross-language similarities in contrast to the French narrator, which introduces less expressive variability in the considered categories. This fact suggests that personal styles exist in storytelling, which was an expected result a priori. Nonetheless, we have found particular acoustic profiles for the three storytelling categories under analysis across narrators beyond such personal styles. Secondly, post-character utterances are more differentiated from the rest of categories, whereas the neutral utterances fall in between descriptive and post-character utterances, with fewer differences when compared to descriptive utterances. This pattern can be clearly observed in the combined-groups plots of the discriminant analyses (see Fig. 1) and in the classification results (see Table 2). Certainly, there is a cross-language similarity in the sense that post-character utterances are clearly discriminated, whereas neutral and descriptive utterances are less distinguishable, showing comparable classification results. Probably, descriptive utterances contain concrete speech subtle nuances (e.g., an extra emphasis in adjectives [2]) that should be studied in detail in the future to find other expressive evidences beyond the ones observed in the current work.

5. Conclusions and Future Work

In this paper, the role of prosody and VoQ across four languages (English, French, German and Spanish) in text-dependent categories of storytelling speech has been explored through several perceptual, statistical and discriminant analyses. Three prosodic features (mean f0, mean intensity and number of silent pauses) and two VoQ parameters (mean HNR and MDQ) explained a relatively similar proportion of the variance among storytelling categories in all considered languages. This finding implies that the analysed narrators use these parameters in relatively similar way to differentiate categories within the same story uttered in their native languages. Moreover, these findings are present in the majority of cases beyond personal styles of the narrators. In line with these similarities, the perceptual test performed to validate neutral corpora led to substantial agreement among raters, thus, highlighting that the narrators used neutral expressiveness for the same sentences across languages.

In the future, we plan to expand the study with a thorough perceptual-acoustic analysis considering more storytelling expressive categories, such as suspense situations. As a long-term goal, we plan to generate synthetic storytelling speech after deriving an acoustic model to be later included within a Text-To-Speech synthesis system.

6. Acknowledgements

Raúl Montaño thanks the support of the European Social Fund (ESF) and the Catalan Government (SUR/DEC) for the predoctoral FI grant No. 2015FI_B2 00110. We also thank the annotators and the people that took the perceptual test for their disinterested help and Marc Freixes for his support when needed.

7. References

[1] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, Saarland University, 2004.

[2] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for storytelling applications," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1137–1144, 2006.

[3] D. Doukhan, A. Rilliard, S. Rosset, M. Adda-Decker, and C. d'Alessandro, "Prosodic analysis of a corpus of tales," in *INTERSPEECH 2011 - 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, 2011, pp. 3129–3132.

[4] J. Adell, A. Bonafonte, and D. Escudero, "Analysis of prosodic features towards modelling of emotional and pragmatic attributes of speech," *Procesamiento de Lenguaje Natural*, vol. 35, pp. 277–283, 2005.

[5] C. O. Alm and R. Sproat, "Perceptions of emotions in expressive storytelling," in *INTERSPEECH 2005 - 6th Annual Conference of the International Speech Communication Association*, Lisbon, Portugal, 2005, pp. 533–536.

[6] W. F. Brewer and E. H. Lichtenstein, "Stories are to entertain: A structural-affect theory of stories," *Journal of Pragmatics*, vol. 6, no. 5-6, pp. 473–486, 1982.

[7] S. Patel, K. R. Scherer, J. Sundberg, and E. Björkner, "Acoustic markers of emotions based on voice physiology," in *Proceedings of the 5th International Conference on Speech Prosody*, Chicago, USA, 2010.

[8] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, no. 4, pp. 417–435, 2009.

[9] P. Liu and M. D. Pell, "Processing emotional prosody in Mandarin Chinese: A cross-language comparison," in *Proceedings of the 7th International Conference on Speech Prosody*, 2014, pp. 95–99.

[10] S. Grawunder and B. Winter, "Acoustic correlates of politeness: prosodic and voice quality measures in polite and informal speech of Korean and German speakers," in *Proceedings of the 5th International Conference on Speech Prosody*, Chicago, USA, 2010.

[11] H. Calsamiglia and A. Tusón, "Los modos de organización del discurso (Chapter 10)," in *Las Cosas del decir: manual de análisis del discurso*. Ariel Lingüística, 1999, pp. 269–323.

[12] N. Mamede and P. Chaleira, "Character identification in children stories," in *Advances in Natural Language Processing*, ser. Lecture Notes in Computer Science, J. L. Vicedo, P. Martínez-Barco, R. Muñoz, and M. Saiz Noeda, Eds. Springer Berlin Heidelberg, 2004, vol. 3230, pp. 82–90.

[13] C. Monzo, F. Alías, I. Iriondo, X. Gonzalvo, and S. Planet, "Discriminating expressive speech styles by voice quality parameterization," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007, pp. 2081–2084.

[14] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, pp. 1170–1179, 2013.

[15] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

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

[17] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193, pp. 97–110, 1993.

[18] K. R. Scherer, "Vocal correlates of emotional arousal and affective disturbance," in *Handbook of Psychophysiology: Emotion and social behavior*, H. Wagner and A. Manstead, Eds. Oxford, UK: Wiley & Sons, 1989.

[19] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta oto-laryngologica*, vol. 90, no. 1-6, pp. 441–451, 1980.

[20] M. Jackson, P. Ladefoged, M. Huffman, and N. Antoñanzas-Barroso, "Measures of spectral tilt," *Journal of the Acoustical Society of America*, vol. 77, no. S1, pp. S86–S86, 1985.

[21] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[22] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter - A new method for quantification of the glottal flow," *Speech Communication*, vol. 22, no. 1, pp. 67–79, 1997.

[23] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 960–964.

[24] C. Drioli, G. Tisato, P. Cosi, and F. Tesser, "Emotions and voice quality: Experiments with sinusoidal modeling," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, Geneva, Switzerland, 2003, pp. 127–132.

[25] J. P. Goldman, "EasyAlign: An automatic phonetic alignment tool under Praat," in *INTERSPEECH 2011 - 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, 2011, pp. 3233–3236.

[26] T. Kisler, F. Schiel, and H. Sloetjes, "Signal processing via web services: the use case WebMAUS," in *Proceedings of the Digital Humanities*, Hamburg, Germany, 2012, pp. 30–34.

[27] B. Bigi and D. Hirst, "SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody," in *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China, 2012.

[28] S. Planet, I. Iriondo, E. Martínez, and J. A. Montero, "TRUE: an online testing platform for multimedia evaluation," in *Programme of the Workshop on Corpora for Research on Emotion and Affect*, 2008, p. 61.

[29] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing emotions in a foreign language," *Journal of Nonverbal Behavior*, vol. 33, no. 2, pp. 107–120, 2009.

[30] B. Bigi and D. Hirst, "Free-marginal multirater Kappa: An alternative to Fleiss' fixed-marginal multirater Kappa," in *Learning and Instruction Symposium*, Joensuu, Finland, 2005.

[31] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[32] R. L. Brennan and D. J. Prediger, "Coefficient kappa: Some uses, misuses, and alternatives," *Educational and psychological measurement*, vol. 41, no. 3, pp. 687–699, 1981.

[33] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[34] B. Andreeva, G. Demenko, B. Möbius, F. Zimmerer, J. Jügler, and M. Oleskowicz-Popiel, "Differences of pitch profiles in Germanic and Slavic languages," in *INTERSPEECH 2014 - 15th Annual Conference of the International Speech Communication Association*, Singapore, 2014, pp. 1307–1311.

[35] IBM, "Home page for PASW/SPSS software [Computer program] (v.22)," http://www.spss.com/software/statistics/.