Synthesizing sports commentaries: One or several emphatic stresses?

Sandrine Brognaux[1,2,3] *Thomas Drugman*[3], *Marco Saerens*[2]

[1]Cental, [2] ICTEAM, Université catholique de Louvain, Belgium
[3]TCTS Lab, Université de Mons, Belgium

sandrine.brognaux@uclouvain.be, thomas.drugman@umons.ac.be, marco.saerens@uclouvain.be

Abstract

Emphatic stresses are known to fulfill essential functions in expressive speech. Their integration in speech synthesis usually relies on a prosodic annotation of the training corpus. Emphasized syllables are then assigned a single label or can receive several labels according to their acoustic realization. While it is more complex to predict those various labels for a new text to synthesize, it might allow for a better rendering of the stress in the synthesized speech. This paper examines whether the use of more than one emphatic label improves the perceived expressivity of the synthesized speech. It relies on a manually-annotated expressive corpus of sports commentaries. Statistical acoustic analyses show that four distinct realizations of emphatic stresses can be distinguished. However, perceptual tests indicate that the integration of this distinction in HMM-based speech synthesis does not lead to a significant improvement in expressivity. This seems to imply that the different acoustic realizations of the stress are not required to be explicitly annotated in the training corpus.

Index Terms: Emphasis, Emphatic stress, Expressive speech, HMM-based speech synthesis, Prosody.

1. Introduction

Recent research in speech synthesis has been targeting the generation of expressive speech [1, 2, 3]. Strategies have been proposed to modify both voice quality and prosody so as to produce the most natural quality of expressivity. Albeit rather unfrequent in neutral speech, emphasis plays a crucial role in the prosody of expressive speech [4, 5]. This phenomenon relates to highly prominent syllables which are sometimes seen as (particularly prominent) 'pitch accents' [6, 7]. Emphatic stresses are known to fulfill various functions like contrasting or highlighting elements and contribute to the liveliness of the message [4]. Their generation in expressive speech is therefore essential. It is even more the case when synthesizing sports commentaries which have been shown to display high rates of emphatic stresses, falling on specific positions like numbers in scores [8, 9].

Several attempts to integrate emphasis have been proposed, both in unit-selection [10, 4, 5, 11] and HMM-based speech synthesis [7, 12]. They usually rely on a prosodic annotation of the corpus in terms of emphatic stresses. The acoustic characteristics of the corresponding syllables are then learned to be reproduced at synthesis stage. While some annotations present various labels associated with different acoustic realizations of the emphatic stress (like ToBI [13]), most studies only use a single label for emphasis [12, 14].

The obvious advantage of using one single label is that it makes it easier to predict it for a new sentence. Most studies investigating the automatic prediction of emphasis from text have, for that matter, considered only one single emphatic label [5, 6, 14]. Conversely, predicting several emphatic labels from text requires a correlation between the labels and specific distinct functions, which is rarely the case.

However, emphatic stress is often regarded as a gathering of different kinds of stresses with various functions, positions, and, importantly, different acoustic realizations (as proposed in ToBI [13]). While some studies have mentioned the potential existence of different *levels* of emphasis [4, 15], we rather believe that different *kinds* of emphasis may co-exist, with no specific order relation between them. If one label is associated with each type of acoustic realization, it allows the training of more acoustically-consistent models, more inclined to generate suitable emphatic stresses in speech synthesis.

The question that still remains is whether the use of a single emphatic label in HMM-based speech synthesis still allows for an appropriate rendering of the various acoustic realizations. In other words, it should be assessed whether the acoustically-different emphatic stresses are learned by the models, based on the linguistic context, or if the distinction requires to be made explicit by annotating with distinct emphatic labels.

This study investigates the benefit of annotating emphatic stresses for expressive HMM-based speech synthesis with several labels instead of one. For that matter, it relies on a large corpus of sports commentaries. The corpus is spontaneous while containing a natural variety of prosody, conversely to studies relying on artificially-produced emphatic stresses by actors (see e.g. [5]). Besides, sports commentaries are characterized by a high density of emphatic stresses with strong acoustic correlates [8], which makes them much more suitable for the study of emphasis than rather neutral read speech as used in [12]. Emphatic stresses were manually annotated in the corpus and were statistically analyzed in order to define several sets of labels, corresponding to stresses with distinct acoustic realizations. The manual annotation having been realized on a functional basis, our study partly answers Hirst's critics [16], i.e. the fact that prosodic function and form tend to be merged in prosody annotation. The objective is here to distinguish between various forms of a single emphatic function and assess the resulting improvement reached in the expressivity of the synthesized speech.

The paper is organized as follows. Section 2 presents the corpus and its emphatic annotation. The statistical acoustic analysis of the emphatic stresses is further described in Section 3. The integration of different sets of emphatic labels in HMM-based synthesis is investigated in Section 4 through a perceptual evaluation. Finally, Section 5 concludes the paper.

10.21437/SpeechProsody.2014-41

2. Corpus design

This study is based on a corpus of live commentaries of two basketball games, uttered by a professional French commentator and recorded in sound-proof conditions. The speaker watched the game and commented it without any prompt. The issue with sports commentaries corpora is usually the high level of background noise which precludes their precise acoustic analysis [17]. Conversely, our corpus exhibits the advantage of being spontaneous and of high acoustic quality, being therefore suited for speech synthesis. Both matches star the Spirou Belgian team with very tight final scores, which induces a high level of excitation. The corpus lasts 162 minutes, silences included.

The corpus was orthographically and phonetically transcribed with [18], with manual check. The phonetic transcription was aligned with the sound using Train&Align [19] and other linguistic information (syllables, parts of speech, etc.) was generated by Elite [20]. Manual annotation of emphatic stresses was realized by assigning a 'F' label to syllables for which an emphatic function was perceived. The annotation results from two or three listenings of each sequence of 4-5 words and was submitted to a second check by the same annotator. In total, 803 syllables were annotated 'F'. Twenty percents of the corpus were annotated by a second expert, and rather high kappa scores were reached (see [8] for the complete analysis of the prosodic annotation).

3. Statistical analysis of emphatic stresses

The statistical acoustic analysis of the emphatic stresses in the corpus consists in four steps. First, a set of acoustic features is extracted for each emphasized syllable (Section 3.1). Dimensionality reduction techniques are then used to minimize the set of features and delete potential redundancy (Section 3.2). The reduced feature set is then used to cluster the emphasized syllables, as an attempt to find the more suitable number of distinct emphatic stresses (Section 3.3). These new sets of stresses are then investigated for potential correlations with specific linguistic contexts (Section 3.4). For further information about the exploited statistical methods, see [21, 22].

3.1. Extraction of acoustic features

For each emphasized syllable, 65 acoustic values are extracted. The first features consist in prosodic measurements: F0 extracted with SRH [23] (mean, max, etc.), energy (mean, max, etc.) and duration (both of syllable and nucleus). A prominence value is added by PromGrad [15] which assigns a prominence score from 0 to 4 to each syllable, on an acoustic basis. Two additional features indicate the presence of a preceding or following silence and its duration. Finally, contextual information, i.e. comparisons with the acoustic values of the two previous and the next syllable, are also computed as they were shown to be efficient for prominence detection in French [24, 15]. These latter measurements are only extracted if both syllables are not separated by a silence, as it is known that silences tend to be associated with a resetting of the prosodic parameters.

It should be noted that, as in [25], duration values are normalized with respect to the average and standard deviation of the duration of the corresponding phonemes. This choice relies on the fact that the nature of the phoneme clearly affects its duration [26, 27]. Missing values (for contextual information) are replaced by the average value of the feature. Finally, all variables are normalized into standard scores.

3.2. Dimensionality reduction

The second stage of our analysis aims at reducing the number of features. For that matter, a principal component analysis (PCA) is carried out on the data. The scree plot (see Figure 1) shows the contribution of the components to the global variance. Since there is no universal technique for selecting the natural number of dimensions, we relied on two popular rules of thumb: (i) keeping the dimensionality accounting for 70% of total variance and (ii) removing the dimensions for which the contribution to the global variance remains stationary. Based on these considerations, we chose to keep ten and four dimensions.

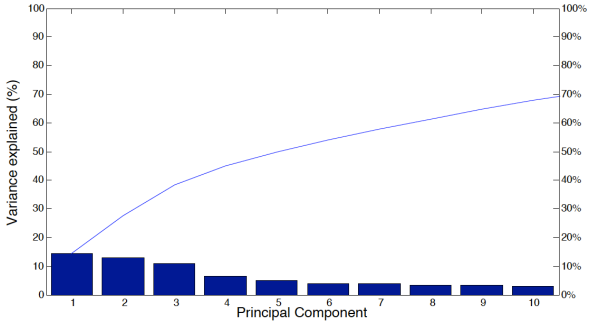

Figure 1: *Scree plot displaying the proportions of global variance carried out by the ten first components of the PCA.*

An insightful analysis regards the weights assigned to the different variables for each component. Interestingly, the first component is clearly related to energy, all 20 higher weights being assigned to energy values. The second component is linked to F0 with the 18 first weights corresponding to F0-based features. Finally, duration-based measurements have most of the heaviest weights in the third PCA component. The fourth dimension is a mix of different types of variables (esp. F0 and energy). The three higher weights for the first three components are assigned respectively to mean energy, mean F0 and syllable duration.

3.3. Clustering

A clustering is now carried out on the reduced data obtained by PCA. The main objective is to define different sets of emphatic stresses characterized by distinct acoustic values.

We first apply a Ward dendrogram [28]. The advantage of this clustering technique it that it visually shows the gathering of the various clusters, which helps in determining the natural number of clusters. Figure 2 shows the dendrogram obtained on the first four PCA components of our data. The algorithm assigns a unique color to each group of nodes where the linkage is less than a specific fixed threshold. This dendrogram clearly shows 4 distinct groups of syllables. Interestingly, when applied to the first 10 PCA components, the dendrogram also points at four distinct clusters.

A second advantage of first applying a dendrogram algorithm is that the centroids of the generated clusters can then be exploited for the initialization of a K-means clustering, which is done in this second stage of our analysis. To assess the quality of the clustering obtained when using various numbers of clusters, we also compute K-means clusterings with 2 to 10 clusters. For that matter, initialization points are selected randomly and the algorithm is run 50 times, the best clustering being kept for

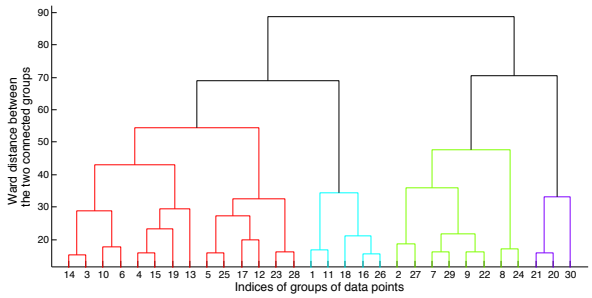

Figure 2: *Ward dendrogram applied to the first four PCA components of our data (the leaves in the plot correspond to more than one data point).*

analysis.

The silhouette value provides an evaluation of the clustering quality, the higher the value, the better the distinction between the various clusters [29]. Figure 3 interestingly shows that, even when launching a K-means with random initialization on the 4 PCA components of our data, the algorithm achieves the best clustering quality with four clusters, which confirms what was shown by the dendrogram. It should be noted that a similar curve, with a peak for 4 clusters, is also observed when launching the K-means on 10 PCA components instead of four. Another interesting finding is that the silhouette value reached with 4 clusters and random initialization is identical to that obtained when initializing the K-means on the centroids of the dendrogram, which might indicate a potential similarity between both clusterings. It should be highlighted, however, that the obtained silhouette values are rather low, indicating a rather uncertain distinction between the clusters.

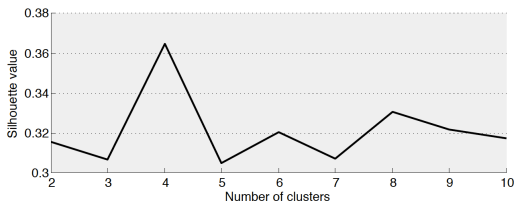

Figure 3: *Value of the silhouette of the K-means on four PCA components according to the number of clusters, with random initialization.*

The rand-index [30] allows comparing two clusterings. Its value ranges between 0 and 1, 1 corresponding to two identical clusterings. Paired comparisons were performed for four K-means clusterings in 4 clusters: with initializations on centroids of dendrogram on 4 (i) and 10 PCA components (ii) and with random initialization on 4 (iii) and 10 PCA components (iv). All rand-index values reach a level above 0.93 which indicates a certain stability, all clusterings converging towards the same solution.

The first of the four clusterings is used in the remainder of this study. A prosodic analysis of the syllables contained in the four clusters indicates that each cluster can be associated to a specific realization in terms of the three main prosodic features (i.e. energy, F0 and duration), as shown in Table 1. To investigate whether the increase in the number of clusters goes in line with an increase in the naturalness of the expressivity, the version with 10 clusters, as obtained with random initialization on the 4 PCA components, is also assessed in the perceptual evaluation. This clustering provides a vectorial quantification of the acoustic space, each region being assigned to a different cluster. It should be noted, however, that it is obviously more complex to predict such a high number of tags from a text to synthesize.

Table 1: *Acoustic characteristics of the clusters, compared to average acoustic values of emphatic stresses.*

Cluster	Energy	F0	Duration
Cluster 1	-	+	-
Cluster 2	-	-	-
Cluster 3	+	+	-
Cluster 4	+	+	+

3.4. Correlation between clusters and linguistic contexts

Potential correlations between the four defined clusters and linguistic information (syllable position, structure, etc.) are investigated. HMM-based speech synthesizers relying on such contextual criteria to cluster the models, correlations would indicate a possible automatic distinction between the various acoustic realizations. This would imply that the clusters would not need to be explicitly distinguished in the annotation.

We analyzed 13 linguistic variables used as contextual information for synthesis: position of the syllable or word in the word or rhythmic group (RG), amount of syllables in word and RG, amount of (content) words in RG, structure of the syllable, nature of the nucleus and part of speech of the word. Seeing the high amount of samples, Chi-square tests tend to be significant for most variables. Cramer's V [31] allows interpreting chi-squares for high effectives. Table 2 shows that only weak associations (i.e. $V < 0.2$ [32]) can be seen between the acoustic clusters and contextual linguistic information.

Table 2: *Correlation between the four clusters and linguistic contextual information (first five variables).*

Variable	Cramer's V
Syllable position in word	0.1379
Nature of the vowel	0.1165
Word position in rhythmic group (forward)	0.1164
Syllable position in word (forward)	0.1141
Syllable position in word (backward)	0.1137

Interestingly, the highest value (i.e. 0.14) is assigned to 'syllable position in word', mainly informative about whether a syllable is initial or final. The omnipresence of both syllable and word position in the ranking drove us to investigate whether some acoustic differences may be due to final syllables at the end of the RG. In that case, they might coincide with what is commonly referred to as boundary tone [13], which could influence their realization. Table 3 shows that the acoustic values of those syllables are significantly higher compared to the other emphatic stresses (respectively p=4.1e-05, p=1.2e-04 and p=1.6e-08 for a bilateral ranksum test performed on 82 final emphatic syllables and 721 other emphatic syllables). This indicates that the distinction between the clusters can partly be explained by linguistic contextual information. Associations

are however rather weak, suggesting that other factors probably play a role in the acoustic realization of emphatic stresses.

Table 3: *Acoustic realizations of final emphatic stresses (end of word at the end of RG) and other emphatic stresses, together with their 95% confidence intervals.*

Emphatic stress	Mean F0 (Hz)	Syllable Dur (z-score)	Mean Energy (dB)
Final	262.9 ±7.4	1.9 ±0.6	49.4 ±1.8
Other	242.9 ±3	0.84 ±0.1	44.2 ±0.4

4. Speech synthesis: A perceptual study

4.1. Evaluation protocol

In order to assess the quality of the expressivity produced when integrating various types of emphatic stresses, several HMM-based speech synthesizers [33] were built, relying on the implementation of the HTS toolkit (version 2.1) publicly available in [34]. For each synthesizer, 90% of the corresponding database was used for the training (called the *training set*), leaving around 10% for the synthesis (called the *synthesis set*). As filter parameterization, we extracted the Mel Generalized Cepstral (MGC) coefficients traditionally used in parametric synthesis. As excitation modeling, the Deterministic plus Stochastic Model (DSM [35]) of the residual signal was used to improve naturalness. Emphatic annotation was used as contextual information, in the same way as linguistic information.

Three models are compared: the baseline model (*Baseline*), using only one emphatic stress, and the models with 4 (*4 Stresses*) and 10 (*10 Stresses*) emphatic stresses, as obtained by annotating the emphatic syllables with the 4 and 10 clusters defined in the previous section. Test sentences were automatically selected from the synthesis set, as being shorter than 5 seconds and displaying at least two emphatic stresses in their annotation. It is indeed much easier to compare short sentences in which more than one difference appears. The test consisted in 18 pairs of sentences, 6 from each comparison, randomly selected from 63 pairs (21 for each comparison).

30 native French-speaking testers, mainly naive listeners, participated in the evaluation. During the test, they could listen to the pair of sentences as many times as wanted. For each comparison, they were first asked whether they heard any difference between both versions of the sentence. If so, they were asked to compare them in terms of naturalness of the expressivity. The scale ranged from -3 (much less natural) to +3 (much more natural). A score of 0 was given if both versions were found to be different but with equivalent naturalness of the expressivity.

4.2. Results

A first interesting finding is that the testers did not hear any difference between both versions for around 20 % of the pairs. This percentage is even higher (i.e. 28%) for models with 4 and 10 clusters which tend to display rather similar intonational patterns. Figure 4 shows the preference percentages for the remaining pairs. Middle sections correspond to pairs which were considered as similarly natural in terms of expressivity. We can observe that the model with 4 emphatic stresses slightly outperforms both other models. This might be explained by the fact that it more accurately synthesizes the various acoustic realizations of the stress. In the10-cluster model, we notice a degradation which may be due to the reduced number of occurrences for each stress, which is partly alleviated with the 4-cluster model. However, the preference for 4 clusters rather than one single stress (i.e. the baseline) is quite weak and is not statistically significant (p=0.11 with a unilateral ranksum test comparing the average percentage of preferences on the 30 testers).

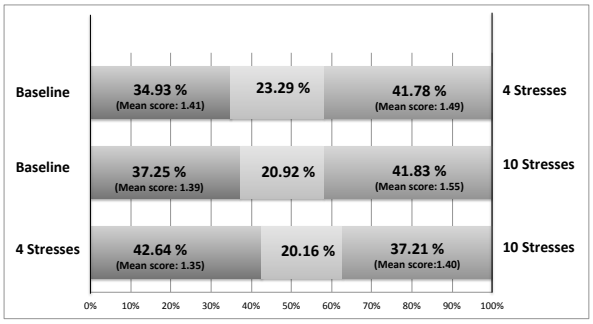

Figure 4: *Percentage of preferences for each model in the three comparison pairs.*

Figure 4 also shows, in parenthesis, the mean score obtained by the three models when preferred in the comparison. These scores are barely higher than 1 because testers mostly assigned a score of '1', reflecting only a 'slight' preference.

5. Conclusion

Emphasis is known to play a crucial role in expressive speech. Its generation in speech synthesis usually relies on a prosodic annotation of the training corpus. For that matter, emphatic stresses can be assigned a single label or be divided into distinct labels according to their acoustic realization. While the prediction of a single label from text is easier, the use of different tags might allow for the generation of more suitable stresses. The question is then whether the use of several emphatic labels effectively improves the naturalness of the expressivity.

The objective of this paper was precisely to answer this latter question by investigating HMM-based speech synthesis using one or several emphatic labels. Statistical acoustic analyses allowed determining 4 and 10 distinct emphatic labels based on a set of extracted acoustic features at the syllable level. The definition of 4 clusters was shown to achieve the best clustering quality. The model with 10 labels was computed to propose a vectorial quantification of the acoustic space. Both models were compared to a baseline model using a single emphatic label.

Perceptual tests showed that the model with four emphatic stresses is slightly preferred over both other models. However, the differences are not significant. While participants did not perceive any difference in 20% of cases, more than 20% of the remaining pairs were scored as 'similar' regarding the naturalness of expressivity. For pairs for which a preference was given, the score was usually low, denoting a weak degree of preference. These results tend to indicate that it might not be required to explicitly annotate different kinds of emphatic stresses in the corpus, when using HMM-based speech synthesis. This is a clear advantage for the annotation of the text to synthesize. However, this finding should be confirmed with further investigations on data with other speaking styles and languages.

6. Acknowledgements

The two first authors are supported by FNRS. The project is partly funded by the Walloon Region Wist 3 SPORTIC.

7. References

[1] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, "Constructing emotional speech synthesizers with limited speech database," in *ICSLP*, 2004, pp. 1185–1188.

[2] J. Yamagishi, K. Onishi, T. Musuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IECE Transactions on Information and Systems*, vol. E88-D(3), pp. 502–509, 2005.

[3] L. Qin, Z.-H. Ling, Y.-J. Wu, B.-F. Zhang, and R.-H. Wang, "HMM-based emotional speech synthesis using average emotion models," in *ISCSLP*, 2006, p. 233240.

[4] R. Fernandez and B. Ramabhadran, "Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis," in *SSW6*, 2007.

[5] V. Strom, R. Clark, and S. King, "Expressive prosody for unit-selection speech synthesis," in *Interspeech*, 2006.

[6] J. Hirschberg, "Accent and discourse in context: Assigning pitch accent in synthetic speech," in *AAAI*, 1990.

[7] L. Badino, J. Andersson, J. Yamagishi, and R. Clark, "Identification of contrast and its emphatic realization in HMM-based speech synthesis," in *Interspeech*, 2009.

[8] S. Brognaux, B. Picart, and T. Drugman, "A new prosody annotation protocol for live sports commentaries," in *Interspeech*, 2013.

[9] B. Picart, S. Brognaux, and T. Drugman, "HMM-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation." in *8th ISCA Speech Synthesis Workshop (SSW8)*, 2013.

[10] A. Raux and A. Black, "A unit selection approach to F0 modeling and its application to emphasis," in *ASRU*, 2003.

[11] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alavrez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Interspeech*, 2007.

[12] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *ICASSP*, 2010.

[13] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *International Conference on Spoken Language Processing (ICSLP)*, 1992, pp. 867–870.

[14] D. Hovy, G. Krishna Anumanchipalli, A. Parlikar, C. Vaughn, A. Lammert, E. Hovy, and A. Black, "Analysis and modeling of "focus" in context," in *Interspeech*, 2013.

[15] J.-P. Goldman, M. Avanzi, A. Auchlin, and A. C. Simon, "A continuous prominence score based on acoustic features," in *Interspeech*, 2012.

[16] D. Hirst, "Form and function in the representation of speech prosody," *Speech Communication*, vol. 46, pp. 334–347, 2005.

[17] J. Trouvain, "Between excitement and triumph - live football commentaries in radio vs. TV," in *17th International Congress of Phonetic Sciences (ICPhS XVII)*, 2011.

[18] J.-P. Goldman, "Easyalign: an automatic phonetic alignment tool under Praat," in *Interspeech*, 2011, pp. 3233–3236.

[19] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort, "Train&Align: A new online tool for automatic phonetic alignments," in *IEEE Workshop on Spoken Language Technologies*, 2012, pp. 410–415. [Online]. Available: http://cental.fltr.ucl.ac.be/train_and_align/

[20] V. Colotte and R. Beaufort, "Linguistic features weighting for a text-to-speech system without prosody model," in *Interspeech*, 2005, pp. 2549–2552.

[21] R. Johnson and D. Wichern, *Applied multivariate statistical analysis, 5th Ed.* Prentice Hall, 2002.

[22] A. Izenman, *Modern multivariate statistical techniques*. Springer, 2008.

[23] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Interspeech*, 2011.

[24] J.-P. Goldman, M. Avanzi, A. Lacheret-Dujour, A. C. Simon, and A. Auchlin, "A methodology for the automatic detection of perceived prominent syllables in spoken French," in *Interspeech*, 2007, pp. 98–101.

[25] S. Brognaux, T. Drugman, and R. Beaufort, "Automatic detection of syntax-based prosody annotation errors," in *IEEE Spoken Language Technology Workshop (SLT)*, 2012.

[26] H. A. Rositzke, "Vowel-length in general American speech," *Language*, vol. 15, pp. 99–109, 1939.

[27] A. Di Cristo, "De la microprosodie à l'intonosyntaxe," Ph.D. dissertation, Université de Provence, Aix-en-Provence, 1985.

[28] J. Ward, "Hierarchical grouping to optimise an objective function," *Journal of the American Statistical Association*, vol. 58, p. 236244, 1963.

[29] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[30] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American statistical Association*, vol. 66, pp. 846–850, 1971.

[31] H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[32] L. M. Rea and R. A. Parker, *Designing and Conducting Survey Research*. Jossey-Bass, 1992.

[33] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51(11), pp. 1039–1064, 2009.

[34] Hmm-based speech synthesis system (hts). [Online]. Available: http://hts.sp.nitech.ac.jp

[35] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20(3), pp. 968–981, 2012.