Automatic extraction of speech rhythm descriptors for speech intelligibility assessment in the context of Head and Neck Cancers

Robin Vaysse, Jérôme Farinas, Corine Astésano, Régine André-Obrecht

Automatic extraction of speech rhythm descriptors for speech intelligibility assessment in the context of Head and Neck Cancers

Robin Vaysse[1,2]*, Jérôme Farinas*[1]*, Corine Astésano*[2]*, Régine André-Obrecht*[1]

[1]IRIT, Université Paul Sabatier, CNRS, Toulouse, France
[2]Octogone-Lordat, Université de Toulouse, UT2J, Toulouse, France

robin.vaysse@irit.fr, jerome.farinas@irit.fr, corine.astesano@univ-tlse2.fr, regine.andre-obrecht}@irit.fr

Abstract

The temporal dimension of speech acoustics is rarely taken into account in automatic models for Speech Intelligibility evaluation, although the rhythmic recurrence of phonemes, syllables and prosodic groups are allegedly good predictors of speech intelligibility. The present study aims at unravelling those automatic parameters that best account for the different levels of the speech signal's rhythmic structure, and to evaluate their correlation with a perceptual intelligibility measure. The parameters are extracted from the Fourier Transform of the amplitude modulation of the signal (Envelope Modulation Spectrum) [1, 2]. A Lasso linear model for feature selection is first implemented to select the most relevant parameters, and a SVR regression analysis is run to reveal the best parameters' combination. Our analyses of EMS, using data from the French corpora of cancer speech C2SI [3], show strong performances of the automatic prediction, with a correlation of 0.70 between our model and an intelligibility evaluation score by speech-pathologists. In particular, the highest correlation with speech intelligibility lies in the ratio between the energy in the low frequency band (0.5-4 Hz that represents slow rhythmic modulations indicative of prosodic groups) and in the higher one (4-10 Hz that represents fast rhythmic modulations like phonemes).

Index Terms: Automatic Speech Processing, speech rhythm modeling, perceptual speech intelligibility, pathological speech

1. Introduction

Radiation and/or surgical procedures following Head and Neck Cancers (H&NC) can have an impact on speech production. They generally result in an alteration of speaker's intelligibility, which affects the patient's quality of life on a daily basis. Scoring and monitoring the evolution of speech intelligibility is an important step in building a therapeutic protocol and ensuring an effective follow-up of speech remediation protocols adapted to patients' lesions. It is thus of importance to find operational measures of speech intelligibility.

Defining speech intelligibility has always been a challenge [4], since it alternatively refers to the performance of telecommunication systems, sound systems or human speech. In this paper, we refer to intelligibility as "listeners' ability to recognize words and/or speech sounds produced by the speaker" [3]. Our goal is to uncover the link between some phonetic-acoustic and perceptual features of speech intelligibility, and to propose automatic metrics both replicable and less prone to variability. Indeed, listeners' judgment exhibits large inter- and intra-listeners' variability [5].

In response to the quest for objectivity, automatic speech processing has proposed various lines of research. Automatic speech recognition systems (ASR) and speaker identification systems (SIS) address speech intelligibility at an acoustic-phonetic level. Intelligibility scores can be correlated to word-error rates of an ASR system [6, 7] or calculated through representative features for speakers' characteristics "i-vectors" or "x-vectors" in SIS [8, 9].

These methods however do not take into account higher-level speech cues to speech intelligibility (such as syllables, prosodic prominence, rhythmic groups coherence), nor the temporal dimension of speech. But these rhythmic features play an important role in speech fluency. H&NC patients may indeed have difficulties reaching some articulatory targets, hence resulting in non-fluent speech (eg. slower speech rate, difficult/slower phoneme and syllable coarticulation). In turn, non-fluent speech leads to disruption of prosodic grouping and sentence structuring, which impairs speech intelligibility.

Because periodicity allows for speech structuring and perceptual prediction [10], rhythm is a major characteristic of speech. Speech rhythm can be defined as the recurrence of strong elements, typically prominent syllables, and their temporal organisation against unstressed syllables [11]. Prominent syllables are perceived as 'beats' delimiting speech units (typically rhythmic groups), while syllables are perceived as 'semi-beats' [12]. Rhythmic groups are also hierarchically organized, revealing the depth of prosodic structuring according to boundary strength. In French, two prosodic levels have been proposed: the Accentual Phrase (AP) and the Intonational Phrase (IP) [13]. Note that IPs may also correspond to breath groups.

Another line of work in automatic speech processing consists in tracking temporal and rhythmic features of speech, through the analysis of intervocalic or interconsonantic duration variations [14, 15, 16, 17]. These methods however require precise manual or automatic annotations of speech segments in order to extract reliable duration measurements. In the context of pathological speech, these may not become available in case of strongly degraded speech.

An alternative to this problem lies in another type of methods relying on the automatic extraction of the signal envelope modulations, or Envelope Modulation Spectrum (EMS), first proposed by [1] and since then used in several studies on speech rhythm [18, 2, 19, 20]). EMS provides spectral analysis of the low-rate amplitude modulations This line of studies is of paramount interest to encompass the whole picture of speech intelligibility. Taking into account the temporal structuring of speech enables to explain the different levels of speech cues that partake in intelligibility, be it at the lowest levels of phonemes or articulatory traits, syllables or the higher levels of prosodic

grouping. On a temporal plane, phonetic segments have a periodicity between 6 and 10 Hz, syllables of roughly 4 to 6 Hz, APs of 2 to 4 Hz and IPs 1 to 2 Hz [21].

The present study wishes to investigate these rhythmic EMS bands with automatic methods in order to account for speech intelligibility in H&NC patients. More specifically, we wish to uncover the potential predictability of the different rhythmic features in the automatic modelling of intelligibility.

2. Extraction of statistical rhythm features

The proposed method used to extract rhythmic parameters of speech signals rely on a few successive steps. The amplitude envelope of the signal is obtained by applying adequate successive Butterworth filters. Empirical step-wise investigation showed that 5 second window frames with a time-lag of 0.5 s are the best configuration for this type of dysfluent speech to extract the relevant envelope signal. A Fourier Transform is applied to the envelope and we extract, from each frame, parameters like the amplitude and frequency of the highest peak or the relative energy on different frequency bands. Follows the computation of basic statistics to characterize their average and their temporal stability across the different 5 second frames. We select the most pertinent statistics through a Lasso regression and use the selected ones with a Support Vector Machine regressor to define an intelligibility score.

2.1. Envelop Modulation Spectrum processing

In order to extract the envelope of the raw signal, [22] proposed a method, adopted in several studies [1, 2]. Two successive Butterworth filters were applied: the first one is a bandpass filter between 700 and 1300 Hz in order to reduce the impact of fundamental frequency and fricative noise [22]; and the second one is a low pass filter with a cutoff at 10 Hz on the magnitudes to constraint the study at the frequency band related to the rhythm. We noticed that the resulting envelope was not correctly correlated to the intensity of the signal. As shown in Figure 1, the envelope of the open vowels like [a] is emphasized compared to the other vowels. An explanation is that the first two formants of [a] are included in this frequency band while only the second (or neither) formant is included for the others. We propose to use a frequency band of 300 - 1000 Hz in order to focus on the first formant of the french vowels. Fig. 1 presents a comparison of the two resulting envelopes.

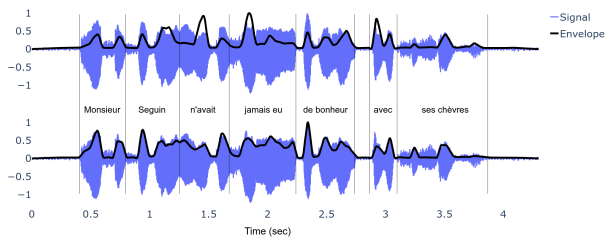

Figure 1: *Comparison between the two filtering methods for amplitude envelope extraction. The upper one uses a band-pass filter of 700-1300 Hz followed by a low-pass filter with cut-off frequency at 10 Hz on the magnitudes of the filtered signal. The second one uses a band-pass filter of 300-1000 Hz and the same low-pass filter*

The following processing is inspired by the study [2]. First a Discrete Fourier Transform (DFT) is applied to each frame of 5s of the envelope, using a Tukey windowing [23], with a time delay of 0.5s between each frame, in order to analyse the periodic information of syllables, words and phrases, and their temporal stability. we choose to limit our analysis to frequencies above 0.5 Hz to avoid periodicity linked to large groups like sentences or larger discourse units. In line with [2], a set of eight parameters are extracted from each EMS frame in order to cover some relevant parts of rhythm relevant rhythmic cues:

- The magnitudes and frequencies of the two most prominent peaks above 0.5 Hz in the EMS, corresponding to the dominant rhythms of the envelope.
- The normalized energy in the 0.5-4 Hz frequency band corresponding to slow variations of the envelope like IPs and APs.
- The normalized energy in the 4-10 Hz frequency band corresponding to faster variations like syllables and phonemes.
- The normalized energy in the 3-6 Hz frequency band, which, according to [24] is a frequency band related to intelligibility.
- The ratio between 0.5-4 and 4-10 Hz which shows which strategy the speaker favors between a good isochrony of syllables and phonemes, and a segmentation of his speech into larger prosodic units.

An example of EMS is illustrated in Figure 2 : three main peaks are emphasized. The first one corresponds to a frequency of 1.97 Hz (patterns of about 500 ms duration) which represents the periodicity of rhythmic groups (APs). The second highest peak at 4.4 Hz (227 ms) shows the duration of syllables, and the last one at 5.8 Hz (172 ms) corresponds to the periodicity of sustained phonemes. The higher the peak of a given frequency F (in Hz), the more patterns of duration $\frac{1}{F}$ (in seconds) are repeated.

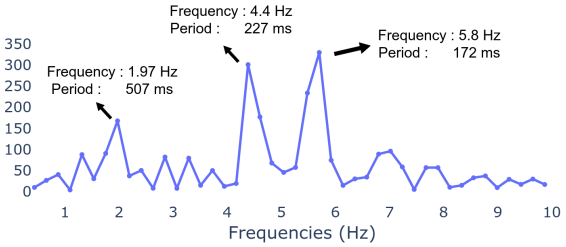

Figure 2: *Example of the Envelope Modulation Spectrum corresponding to the signal in Figure 1*

To characterize these 8 parameters across the 5 seconds sliding windows, we compute their average, standard deviation, skewness and kurtosis, resulting in 32 features to represent the studied speech signal.

2.2. Selection of pertinent features and intelligibity score proposition

As said in the introduction, our purpose is to propose a score, function of these features, which correctly predicts the perceptual intelligibility measure. We worked in two stages to achieve this result. The first step consisted in working on features' selection using the Lasso method [25]. This allowed us to reduce

the number of features and yield better qualitative interpretation on relevant rhythmic features.

In a second step, with these selected set of features, we produced a regression score using a Support Vector machine Regressor (SVR) [26, 27]. This method performs well with small datasets and has already been successfully tested on this kind of task [9]. It consists in finding a linear combination of the features, which values differ by at most a predefined ϵ deviation from target reference values, for all the training data. When this is not feasible, trade-off and slack variables are introduced to solve the optimization problem [28]. In our case, the target reference values are the perceptual intelligibility scores, and the function gives its prediction on the basis of the extracted features.

3. Experimental protocol and evaluation

3.1. Corpus description

The current study is based on the French H&NC speech corpus C2SI [3]. This corpus includes patients suffering from oral cavity or oropharyngeal cancer and healthy speakers, performing different tasks. We used the reading task, where speakers read the first paragraph of the french text *la chèvre de Monsieur Seguin*, a tale by Alphonse Daudet. Our test corpus is thus semantically homogeneous throughout speakers. A total of 105 speakers (24 controls, 81 patients) were used in this study.
A set of 6 different health professionals were asked to assess the intelligibility of each speaker from 0 (unintelligible) to 10 (high intelligibility). The final score for one speaker is computed as the mean of the 6 scores given by the experts.

3.2. Experimental protocol

First of all, features are normalized by subtracting their mean and dividing by their standard deviation across all speakers. For the feature selection process described in Section 2.2, Lasso regression uses an α variable which determines the level of regularization on the features' coefficients: the higher α, the more features will be remove. Several tests with α values belonging to the interval $[0.01, 1]$ are performed to choose the one that yields the best performance of our global system. Regarding the prediction of the intelligibility score, the SVR is implemented using the scikit-learn python toolkit [29]. In order to give the more reliable results on our speech corpora which is relatively small (105 recordings) in the context of machine learning, we use a Leave One Out methodology. The Leave One Out consists of using all the recordings with their perceptual intelligibility scores minus one, in order to train the SVR model and then test the model on the excluded recording. We then repeat this procedure with each recording in order to test the performances of our model on all speakers. The procedure was done using a linear kernel on the SVR; the regularisation parameter C was set to 4 and the ϵ deviation tolerance was set to 0.01 (see [28] for detailed information on these parameters).

To evaluate the relevance of the obtained predicted intelligibility values, two metrics are used, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (1)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (2)$$

Where N is the number of speakers, y_i the perceptual intelligibility measurement of the i-th speaker and $\hat{y}_i$ his/her predicted intelligibility.

To assess the relevance of the feature selection, a sequence of 10 experiments were performed with the SVR approach, one for α values between 0.01 and 1. Another experiment was run on the global set of 32 features.

3.3. Relevance of the selected features

To unravel the contribution of each parameters selected by the Lasso method, we can look at the generated coefficients. Due to the introduction of the parameter normalisation, the higher a coefficient of a parameter, the more it contributes to predict the target value (intelligibility).

The best results in terms of RMSE and MAE for the global system is obtained for $\alpha = 0.1$; this reduced the number of parameters from 32 to 5. The selected parameters with their Lasso coefficient and Pearson correlation with perceptual intelligibility are listed in Table 1.

Table 1: *List of selected features with their associated coefficient generated by the Lasso regression and the Pearson correlation of the feature with the perceptual intelligibility*

Feature	**Linear regression coefficient**	**Pearson Correlation**
Mean of ratio 0.5-4 Hz energy over 4-10 Hz energy	-0.52	-0.65
Skewness of frequency from most prominent EMS peak	0.10	0.3
Mean frequency of 2nd most prominent EMS peak	0.06	0.52
Mean of 0.5-4 Hz energy	0.02	-0.08
Kurtosis of 3-6 Hz energy	0.01	-0.31

It appears that the most important parameter is the mean ratio between low frequency (0.5-4 Hz) and higher frequency (4-10 Hz) energy. This parameter is the one with the best correlation with perceptual intelligibility (Pearson correlation = -0.65). This high negative correlation means that the higher the ratio, the more degraded perceptual intelligibility is. Other important features like standard deviation of the ratio and the mean energy in the 4-10 Hz band are removed by Lasso regression because of their high correlation with the mean of the ratio.

The second best parameter is the skewness of the most prominent EMS frequency peak. We have verified that for all speakers, this skewness is positive so the distribution shape remains quite similar with a majority of low frequencies and a minority of higher ones, synonymous to median and mode values lower than the average. The ampler the skewness, the more important the mode-average delay is.
The third selected parameter is the mean frequency of the second most prominent peak in the EMS. As this parameter shows a strong positive correlation with perceptual intelligibility, this

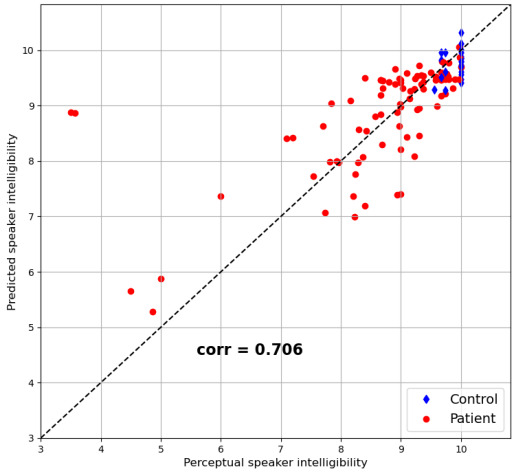

Figure 3: *Automatically predicted intelligibility computed per speaker according to perceptual evaluation.*

means that speakers with good intelligibility have a relative high frequency for the second most prominent frequency. A major issue with this feature is that it can be influenced by the speech rate. We will focus on this particular problem in future works (see 4).

3.4. Validation of the predicted intelligibility score

Table 2: *Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) from the predicted intelligibility using Support Vector Machine regression with and without feature selection*

	RMSE	MAE
SVR with all features	0.91	0.56
SVR with selected features	**0.83**	**0.51**

As said in 3.2, many experiments are performed to find the good association between the selected parameters and the SVR. Only the results of two configurations are reported in Table 2, the worst one with all the 32 parameters and the best one with the 5 selected features : it turns out that this selection marginally improves the quality of our prediction, with RMSE went from 0.91 to 0.83.

To confirm the adequacy between perceptual and predicted intelligibility measures, Figure 3 depicts our predicted intelligibility score estimated by the SVR compared to the reference perceptual evaluation from speech-therapists. We observe that the automatic approach gives a good approximation of intelligibility by only using 5 features extracted from the EMS spectrum. Indeed, we have a strong Pearson correlation of 0.706 between our predictions and the reference values.

However, we can notice that for the two speakers with the least intelligibility, our predictions result in an incorrect estimation. Indeed, both of them have a perceptual intelligibility of 3.5 while our estimations are approximately of 9. The reason behind the misplacement of these speakers is that they have a low value for the mean of the energy ratio compared to other speakers with low intelligibility. Usually, speakers with good intelligibility have a low ratio because their energy between 0.5 to 4 Hz and between 4 to 10 Hz are both high, indicative of regular periodicity at low rhythmic levels (phonemes, syllables) and high rhythmic levels (APs, IPs). On the contrary, for unintelligible speakers, the low frequency energy is strong while the higher is weak. For those two speakers, 4-10 Hz energy is low, but 0.5 to 4 Hz energy is also low, so it results in a small ratio. After listening to their recordings, they show poor syllable and phoneme articulation and they also produce irregular IP and AP duration. Further work needs to be done on this issue using corpora containing more numerous speakers with poor intelligibility in order to compensate for the imbalanced intelligibility distribution.

4. Conclusion and future work

This study aimed at investigating the feasibility of automatic prediction on speech intelligibility scores on pathological speech. We aimed at showing that rhythmic features above the phonetic segment are important predictors of perceived speech intelligibility, especially on the acoustically degraded speech of H&NC patients. Our results confirm that EMS is a particularly relevant analysis, as it allows for investigating all levels of speech units, from phonemes to large prosodic phrases. Our results show an interestingly strong correlation with intelligibility (0.70 Pearson correlation) using a fully automatic method (without manual annotations). Our method is based on rhythmic features, that are usually marginally explored in this type of research. The choice of the Lasso method also allowed us to reduce rhythmic predictors to 5 relevant rhythmic features, allowing to produce global interpretations on the rhythmic characteristics of the speakers. Typically, the ratio between the energy bands ($[0.5, 4]$ Hz: large prosodic units, IPs) and ($[4, 10]$ Hz: syllables and phonemes) is relevant to discriminate rhythmic characteristics of the speech signal.

These results are promising and validate our choice of rhythmic features for pathological speech. Further analyses on the distribution of each speaker will be carried out to refine the interpretation of this ratio: we hypothesize that the strong relevance of this ratio may be indicative of compensatory phenomena of the pathological speakers. Our next step is to take into account speaking rate variability, and to test more precise EMS frequency bands. It will allow us to better characterise the rhythmic levels responsible for the intelligibility loss of our H&NC speakers. Another important research path is to test our model on the spontaneous speech tasks available in the corpus. In the long term, we would like to merge our model with others based on different speech levels, such as those based on i-vectors or x-vectors.

5. Acknowledgements

This work has been carried out thanks to the French National Research Agency in 2018 as part of the RUGBI 2018 project[1] entitled "Looking for Relevant linguistic Units to improve the intelliGiBIlity measurement of speech production disorders" (Grant No. ANR-18-CE45-0008).

[1] https://www.irit.fr/rugbi/

6. References

[1] S. Tilsen and K. Johnson, "Low-frequency Fourier analysis of speech rhythm," *The Journal of the Acoustical Society of America*, vol. 124, pp. EL34–9, Sep. 2008.

[2] Liss Julie M., LeGendre Sue, and Lotto Andrew J., "Discriminating Dysarthria Type From Envelope Modulation Spectra," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 5, pp. 1246–1255, Oct. 2010, publisher: American Speech-Language-Hearing Association. [Online]. Available: https://pubs.asha.org/doi/10.1044/1092-4388(2010/09-0121)

[3] V. Woisard, C. Astésano, M. Balaguer, J. Farinas, C. Fredouille, P. Gaillard, A. Ghio, L. Giusti, I. Laaridh, M. Lalain, B. Lepage, J. Mauclair, O. Nocaudie, J. Pinquier, G. Pouchoulin, M. Puech, D. Robert, and V. Roger, "C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers," *Language Resources and Evaluation*, Jun 2020. [Online]. Available: https://doi.org/10.1007/s10579-020-09496-3

[4] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 95. no. 1, pp. 90–119, 1947.

[5] M. McHenry, "An exploration of listener variability in intelligibility judgments," *American Journal of Speech-Language Pathology*, vol. 20, no. 2, pp. 119–123, 2011. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/1058-0360\%282010/10-0059\%29

[6] L. Fontan, I. Ferrané, J. Farinas, J. Pinquier, J. Tardieu, C. Magnen, P. Gaillard, X. Aumont, and C. Füllgrabe, "Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 9, pp. 2394–2405, 2017. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/2017_JSLHR-S-16-0269

[7] M. Windrich, A. Maier, R. Kohler, E. Noeth, E. Nkenke, U. Eysholdt, and M. Schuster, "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma," *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, vol. 60, pp. 151–6, 03 2008.

[8] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, "Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer," in *Interspeech 2020*, ISCA (International Speech Communication Association). Shangai (fully virtual conference), China: ISCA, Oct. 2020, pp. 4976–4980. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03122735

[9] I. Laaridh, C. Fredouille, A. Ghio, M. Lalain, and V. Woisard, "Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers," in *Proc. Interspeech 2018*, 2018, pp. 2943–2947. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1266

[10] F. Cummins, *Rhythm and Speech*. John Wiley & Sons, 2015, ch. 8, pp. 158–177. [Online]. Available: https://doi.org/10.1002/9781118584156.ch8

[11] A. Di Cristo, "Le cadre accentuel du français contemporain : essai de modélisation. première partie," *Langues (Montrouge)*, 1999.

[12] E. Selkirk, "On the major class features and syllable theory," *Language Sound Structure*, 1984. [Online]. Available: https://ci.nii.ac.jp/naid/10024374723/en/

[13] S.-A. Jun and F. Cecile, "A phonological model of french intonation," *Intonation: Analysis, Modelling and Technology*, 01 2000.

[14] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, pp. 265–292, 06 1999.

[15] E. Grabe and E. Low, "Durational variability in speech and the rhythm class hypothesis," in *Papers in Laboratory Phonology*, Jan. 2002, vol. Vol. 7, pp. 515–546, journal Abbreviation: Papers in Laboratory Phonology.

[16] P. Wagner, "A time-delay approach to speech rhythm visualization, modeling and measurement," in *Prosodic Universals comparative studies in rhythmic modeling and rhythm typology*, 01 2010, pp. 117–146.

[17] J. M. Liss, L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer, and J. N. Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1334–1352, 2009. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/1092-4388\%282009/08-0208\%29

[18] S. Tilsen and A. Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages," *The Journal of the Acoustical Society of America*, vol. 134, pp. 628–39, Jul. 2013.

[19] L. Varnet, M. Ortiz-Barajas, R. Guevara Erra, J. Gervain, and C. Lorenzi, "A cross-linguistic study of speech modulation spectra," *The Journal of the Acoustical Society of America*, vol. 141, pp. 3701–3702, May 2017.

[20] V. Leong and U. Goswami, "Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia," *Frontiers in Human Neuroscience*, vol. 8, 2014, publisher: Frontiers. [Online]. Available: 10.3389/fnhum.2014.00096

[21] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature neuroscience*, vol. 15, no. 4, p. 511, 2012.

[22] F. Cummins and R. Port, "Rhythmic constraints on stress timing in English," *Journal of Phonetics*, vol. 26, no. 2, pp. 145–171, Apr. 1998. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0095447098900705

[23] P. Bloomfield, *Fourier analysis of time series: an introduction*. John Wiley & Sons, 2004.

[24] T. Houtgast and H. J. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of the Acoustical Society of America*, vol. 77, pp. 1069–1077, 1985.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[26] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, 1999, pp. 61–74.

[27] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[28] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. Jordan, and T. Petsche, Eds., vol. 9. MIT Press, 1997. [Online]. Available: https://tinyurl.com/vsbm5433

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.