Rhythmic Proximity Between Natives And Learners Of French Evaluation of a metric based on the CEFC corpus

Sylvain Coulange, Solange Rossato

Rhythmic Proximity Between Natives And Learners Of French Evaluation of a metric based on the CEFC corpus

Sylvain Coulange[a], Solange Rossato[a,b]
[a] LIDILEM, [b] LIG
Université Grenoble Alpes
{sylvain.coulange, solange.rossato}@univ-grenoble-alpes.fr

Abstract

This work aims to better understand the role of rhythm in foreign accent, and its modelling. We made a model of rhythm in French taking into account its variability, thanks to the Corpus pour l'Étude du Français Contemporain (CEFC), which contains up to 300 hours of speech of a wide variety of speaker profiles and situations. 16 parameters were computed, each of them being based on segment duration, such as voicing and intersyllabic timing. All the parameters are fully automatically detected from signal, without ASR or transcription. A gaussian mixture model was trained on 1,340 native speakers of French; any 30-second minimum speech may be computed to get the probability of its belonging to this model. We tested it with 146 test native speakers (NS), 37 non-native speakers (NNS) from the same corpus, and 29 non-native Japanese learners of French (JpNNS) from an independent corpus. The probability of NNS having inferior log-likelihood to NS was only a tendency (p=.067), maybe due to the heterogeneity of French proficiency of the speakers; but a much bigger probability was obtained for JpNNS (p<.0001), where all speakers were A2 level. Eta-squared test showed that most efficient parameters were intersyllabic mean duration and variation coefficient, along with speech rate for NNS; and speech rate and phonation ratio for JpNNS.

Keywords: Rhythm modelling, Speech fluency, Foreign accent, Spoken language processing, L2 pronunciation, Variability of speech

1. Introduction

Perception of foreign accent is mainly due to a difference of pronunciation between what a speaker said and a norm shared by natives of a target language (Alazard, 2013). This difference has been mostly described on the segmental level, through theories of language acquisition, but the role of prosody was proved only recently in accent perception studies (De Meo et al., 2012; Pellegrino, 2012). Today, it is clear that both segmental and suprasegmental levels are important in perception of this difference by native speakers.

Among prosodic parameters, rhythm is one that varies noticeably from one language to another. (Gibbon and Gut, 2001) define rhythm as the recurrence of patterns of weak and strong elements in time. Those elements can be short or long syllables, low or high intonation on vowel or consonant segments. (Di Cristo and Hirst,) define it as the temporal organisation of prominences. (Arvaniti, 2009) talks about perception of repetition and similarity patterns. (Alazard, 2013) refers to physical and psychobiological phenomena like dance and music rhythm, or cardiac rhythm. All these definitions put forward the idea of timing patterns; length, height or intensity patterns in time.

Many studies try to classify languages based exclusively on rhythmic parameters. Most of them rely on vowels and consonant duration, and therefore need an aligned transcription. Furthermore, some recent studies showed similar results with voicing and unvoicing duration or syllables duration (Fourcin and Dellwo, 2013; Dellwo and Fourcin, 2013; Dellwo et al., 2015), parameters that may be automatically detected from signal without any transcription. This is great news since automatic transcription might be an issue with non-native speech, and manual transcription would be too long and costly.

Most studies also show limits of their results due to small amount of speakers or elicitation bias (Fourcin and Dellwo, 2013; White and Mattys, 2007b; Ramus et al., 1999; Gibbon and Gut, 2001; Grabe and Low, 2002). Hence it is necessary to use a large corpus with the largest variation possible, with a lot of spontaneous speech, since rhythm might strongly vary depending on situations and speakers, and even more in spontaneous conversations (Bhat et al., 2010). It is still today extremely important to be able to measure language in its variety, without focusing on very specific conditions (Astésano, 2016).

In this study we tried to model rhythm of French through the recently published Corpus d'Étude pour le Français Contemporain (CEFC, Benzitoun et al. (2016)), which offers a wide variety of French. In order to model this variety as much as possible, we trained a gaussian mixture model on several acoustic parameters, and evaluated the model with test native speakers, and non-native from CEFC and an other corpus. Acoustic parameters are all based on syllabic nuclei detected by a Praat script from De Jong and Wempe (2009), and voicing detection tool from Praat (Boersma and Weenink, 2019). Meanwhile, we also evaluated the efficiency of each parameter to distinguish native and non-native depending on the corpus.

Our concern is to determine which rhythmic parameters make the difference – and are mostly responsi-

ble for the foreign accent – depending on the mother tongue of the speaker. Once these parameters are known, we will be able to offer more appropriate remedial measures to learners.

2. Corpus

As mentioned above, the CEFC corpus offers a great diversity of speech, through numerous situations of conversation and speaker profiles, and a very large amount of speech (about 4 million words). All recordings are also transcribed and aligned, which allowed us to localize the speech of each speaker through conversations. The CEFC also includes about 50 non-native speakers that we used as a non-native test partition for our model. This corpus is a concatenation of 13 subcorpora, some of them already well known like Valibel (Dister et al., 2009) or Clapi[1], and covers variation in Belgium, Switzerland and all regions of France in 2,587 speakers. Speaking situations vary from private conversations to business meetings, through family dinners, radio and television debates, readings of traditional tales or even daily conversation in shops.

Situations are mainly dialogues (481 recordings out of 900), others imply more than 2 speakers (277) or are monologal (144). Most of recordings are face-to-face conversations, but there are also 38 telephone talks, 115 public speeches, 19 recordings from TV and 12 from radio.

In this corpus, women form the majority of the speakers with 1,373 speakers, versus 1,048 men and 166 whom gender isn't provided in the metadata. For once, women are well represented here. Most speakers are between 21 and 60 years old (43.3%), 6.7% are more than 61, 5.3% are from 16 to 20, and 1.2% are less than 15. No age data are given for 38.7% and 4.7% had an age data format non harmonised with the rest of the corpus and aren't taken into consideration here. More information about speakers and situations is given in Coulange (2019).

It is well known today that rhythm can vary a lot depending on regions, situations and speakers (Gibbon and Gut, 2001), the CEFC allows us to consider a large coverage of variation of French for our modelling. However, we have no information about the level of French of non-native speakers, and it might be quite heterogeneous as we find students as well as employees among them. This is why we decided to include 29 non-native speakers with known global and oral level of French, and with quite homogeneous profiles (Japanese university students, same mother tongue, same class).

3. Modeling

3.1. Acoustic parameters

We computed 16 acoustic parameters based on studies of language classification by rhythm (White and Mattys, 2007a; Fourcin and Dellwo, 2013; Pettorino et al., 2013; Rossato et al., 2018, among others), as well as previous studies on foreign accent perception (Bhat et al., 2010; Fontan et al., 2018). Chosen parameters are:

- speech rate (nb. of syllables/segment duration);
- phonation ratio (phonation time/segment duration);
- voiced intervals mean duration, standard deviation and variation coefficient;
- unvoiced intervals mean duration, standard deviation and variation coefficient;
- voiced + unvoiced intervals mean duration, standard deviation and variation coefficient;
- voiced intervals' raw and normalized pairwise comparison index (PVI);
- intersyllabic intervals mean duration, standard deviation and variation coefficient.

All measurements were made on segments of at least 30 seconds of concatenated speech of a single speaker. This is long enough to get reliable values of duration, without being too influenced by local hesitations, laughs or bad measurements. If a speaker speaks less than 30s. in the whole recording, his or her voice is not used (658 speakers in this case). This way, we only keep speakers with at least one speech segment.

3.2. Data partitioning

As we needed at least 30s. of speech for each speaker, and to know if he/she is native or not, we ended up with a training set of 1,340 native speakers (16,884 segments), and 3 sets of test: 1. native speakers test set of 146 speakers (1,919 segments), 2. CEFC non-native speakers test set of 37 speakers (268 segments) and 3. A2-level Japanese learners of French of 29 speakers (96 segments). Table 1 sums up this partitioning.

Set	Training	Test	Test	Test
French status	native	native	non-native	non-native
Corpus	CEFC	CEFC	CEFC	Jp learners
#speakers	1,340	146	37	29
#segments	16,884	1,919	268	96

Table 1: Data partitioning map

3.3. Gaussian Mixture Model

According to Ferrer et al. (2015), gaussian mixture models (GMM) are suitable to model variability. A GMM is a density of probability made from a sum of weighted gaussians. This sum generates a function that fits at best the data distribution. Training the model amounts to find the best parameters of these gaussians to represent the data, through an Esperance-Maximization algorithm. The probability of a vector $\vec{x}$ given a GMM of parameters $\{w_k, \vec{\mu}_k, \Sigma_k\}_{k=1}^{K}$ is then:

$$p(\vec{x}) = \sum_{i=1}^{K} w_i \mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k) \quad (1)$$

[1] Clapi : http://clapi.icar.cnrs.fr

where K is the number of gaussians, w_k is the weight of the gaussian k, such as $\sum_{k=1}^{K} w_k = 1$, and $\mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k)$ the normal function of $\vec{x}$ with mean $\vec{\mu}$ and covariance Σ of k. We used a diagonal covariance to lighten the training, even if some acoustic parameters are correlated to each other. We also chose to limit our GMM to 1,024 components.
The training was implemented in Python using the SciKitLearn Gaussian Mixture library[2].
To get the proximity of a speaker X to our model, we compute the product of the likelihood of the model for each of his segments of speech x:

$$p(X) = \prod_{n=1}^{N} p(\vec{x}_n) \tag{2}$$

To simplify this computation, we turned the product into a sum with the log-likelihood $log\ p(X)$, and we normalized it by the number of segments for each speaker, as it may vary a lot. This eventually gives:

$$log\ p(X) = \frac{1}{N} \sum_{n=1}^{N} log\ p(\vec{x}_n) \tag{3}$$

We computed this mean log-likelihood for each speakers of the native speakers test partition, and compared them to those of non-native speakers with a Wilcoxon-Mann-Whitney test.

4. Results

4.1. Rhythmic proximity depending on speakers

We first compared rhythmic proximity scores of CEFC native speakers (NS) test set and CEFC non-native speakers (NNS) test set. It appeared that the probability of NNS' scores being inferior to NS' didn't exceed the tendency ($p = .067$). It might be due to the heterogeneity of NNS profiles. Indeed, more than half of them are students while the others have various occupations, this is only a hint that makes us think that their level of French and duration of stay in a francophone environment might vary a lot. Mother tongues might also vary significantly even if it isn't explicitly mentioned in the metadata of the corpus – we know only that the 37 speakers come from at least 18 different countries. We know that mother tongue(s)' rhythm may noticeably influence other languages acquisition, as well as duration of stay in the target language speaking countries (Piske et al., 2001; Flege, 1988).
We then compared scores of NS and those of Japanese non-native speakers (JpNNS). This time, the difference is much more significant ($p < .0001$). The rhythmic gap between natives and non-natives is clear here.
Figure 1 present a projection of these scores zoomed on scores superior to -50 (that being 96.6% of NS, 94.6% of NNS and 79.3% of JpNNS). We find that no JpNNS get a score higher than 22.48 and most of them are between 0 and 10, while NNS and NS are mostly between 20 and 40. Means of each population are respectively .74, 21.48 and 25.0 when extremely low results below -50 are ignored. The reasons for these low scores are mainly bad detection of voicing and syllable nuclei due to the voices not being loud enough.
In this figure we also plotted rhythmic score of the French native teacher of the class, whose voice was also present inside some recordings. He got a score of 19.96.
To be sure that our model actually models linguistic rhythm of French, we also compared randomly partitioned NS data. Through 10 comparisons, none of them showed a significant difference.

[2]https://scikit-learn.org

4.2. Correlation between rhythmic score and language proficiency

Along with Japanese students recordings, we have at our disposal 3 different grades for each student: grade for a DELF-type semester global evaluation, based on 4 abilities (oral and written comprehension and expression), as well as the grade obtained for oral expression, and the number of points specifically obtained for their speech fluency. Fluency is evaluated for the oral expression part, and oral expression grade amounts to $\frac{1}{4}$ of the global exam grade. Recordings that we use as a test set actually are those of oral expression part of the exam.
We computed the correlation between the 3 different grades and rhythmic scores for each students. Global exam grades and rhythmic scores showed a strong positive correlation ($r = .598$, $p < .005$; $r^2 = .358$, $p < .005$ and $\rho = .478$, $p < .05$). Students are plotted on Figure 2 depending on their exam grade (x) and rhythmic score (y).
Correlations between scores and grades for oral comprehension and fluency weren't significant though. Grades are very close to each other, and there is only a little difference between students (oral comprehension vary from 17 to 24, fluency from 3 to 5 points). We would need to repeat this analysis with a wider range of grades to ascertain whether we could obtain a significant correlation. Table 2 shows Spearman's correlation coefficient (r), the determination coefficient (r^2) and Pearson's coefficient (ρ) for each type of grade: global, oral expression and fluency; along with associated p-values.

	Global	**Oral**	**Fluency**
r	.598 (p = .003)	.257 (p = .237)	.410 (p = .052)
r^2	.358 (p = .003)	.066 (p = .237)	.168 (p = .052)
ρ	.478 (p = .021)	.315 (p = .144)	.228 (p = .295)

Table 2: Correlation tests results between rhythmic scores and global grade (left), oral expression grade (middle) and fluency (right)

4.3. Efficiency of acoustic parameters

Efficiency of each acoustic parameters to distinguish natives and non-natives was computed through a comparison of NS and NNS on one hand, and of NS and

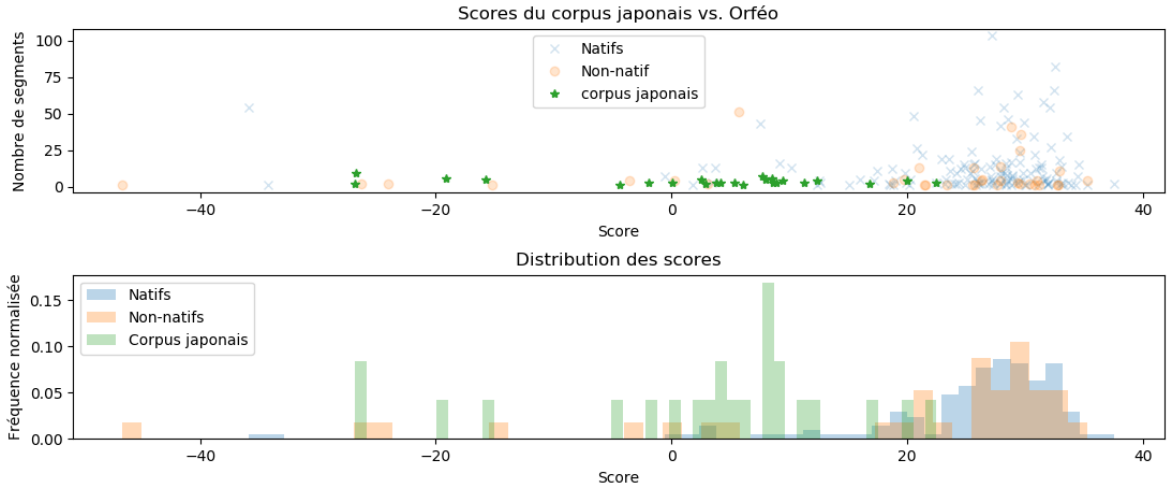

Figure 1: Scores projection of Japanese learners (green), NS (blue) and NNS (orange) superior to -50

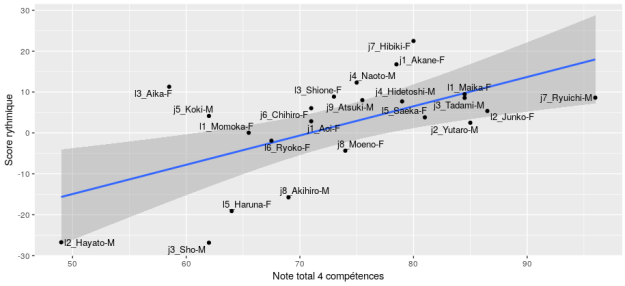

Figure 2: Rhythmic scores in function of global exam grades

Parameter	η^2	P-value		mean η^2	sd η^2
SR	.745	$3.07e^{-58}$	<.0001	.705	.038
$pcdV$	.673	$5.63e^{-48}$	<.0001	.647	.039
$VarcoV$	.415	$7.23e^{-24}$	<.0001	.391	.028
σdV	.373	$4.79e^{-21}$	<.0001	.359	.029
$nPVIdV$	.360	$3.81e^{-20}$	<.0001	.339	.030
$rPVIdV$	.349	$1.99e^{-19}$	<.0001	.327	.035
μdV	.281	$2.64e^{-15}$	<.0001	.246	.033
μdU	.125	$4.78e^{-07}$	<.0001	.123	.003
μdVU	.118	$1.11e^{-06}$	<.0001	.116	.002
σdU	.085	$3.94e^{-05}$	<.0001	.084	.002
$\mu\Delta t$	.084	$4.37e^{-05}$	<.0001	.066	.017
σdVU	.082	$5.46e^{-05}$	<.0001	.081	.001
$Varco\Delta t$	.046	.003	<.01	.029	.015
$VarcoU$	.012	.124	>.05	.012	.010
$\sigma\Delta t$	.011	.157	>.05	.006	.004
$VarcoVU$	.008	.206	>.05	.008	.008

Table 3: η^2 and its p-value between NS and JpNNS for the first sampling iteration, and mean and standard deviation of η^2 on the 3 iterations

JpNNS on the other hand. We will present here only results concerning JpNNS.

Given the small number of non-native speakers' speech segments (96), we resampled the native speakers' segments to 96. Furthermore, to avoid resampling bias, we made 3 different samplings to check if big differences occur.

Let's say first that, by examining values of each acoustic parameter for the first sample, the biggest difference between JpNNS and NS is speech rate (respectively 1.3 syllable per second against 4.0: JpNNS speak much slower than natives), and phonation ratio (15% against 57%: mostly due to frequent silences in non-native speech). Also, we see that voicing interval mean duration is shorter for JpNNS (130 ms) than NS (190 ms), and vary less (standard deviation of 90 ms against 180 ms for natives).

Table 3 presents eta-squared (η^2) of the first sampling, along with their p-value, as well as mean and standard deviation of η^2 through the 3 different samplings. The η^2 value refers to the proportion of variance of the parameter explained by the nativity variable[3]. All parameters until the 12^{th} show a significant difference (<.0001) between NS and JpNNS. We see that speech rate (SR) is explained by 75% by nativity, closely followed by the percentage of phonation ($pcdV$, 67%). Then come metrics implying voicing interval duration: coefficient of variation of voiced intervals ($VarcoV$), standard deviation and mean of voiced intervals duration (σdV, μdV), as well as its normal and row pairwise comparisons ($nPVI\_dV$, $rPVI\_dV$). All of these are explained from 28 to 42% by nativity.

Parameters that compute unvoiced interval duration turned out to be not so efficient to distinguish native and non-native Japanese students (mean of unvoiced intervals duration μdU (13%), mean of voiced followed by unvoiced interval duration μdVU (12%), as well as their respective standard deviation 9 and 8%). Surprisingly parameters computing intersyllabic duration weren't efficient either, only the mean of this duration was significant ($\mu\Delta t$), explained to 8% by nativity. Coefficient of variation of intersyllabic duration, unvoiced and voiced+unvoiced interval duration ($Varco\Delta t$, $VarcoU$, $VarcoVU$) as well as standard deviation of intersyllabic duration ($\sigma\Delta t$) weren't significant in distinguishing natives from Japanese speakers.

[3]In this work, it refers to the fact that a speaker is native or not.

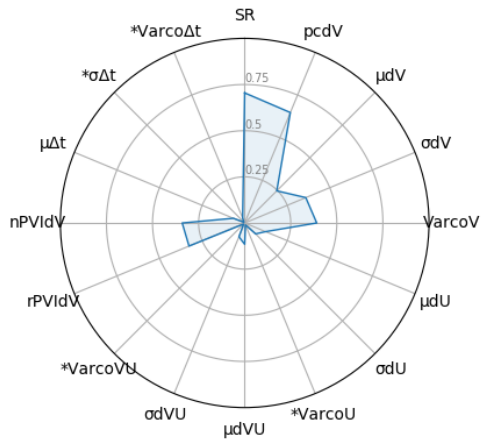

Figure 3: Radar view of η^2 from the first sampling (stared parameters weren't significant for at least one of the 3 resamplings)

Figure 3 is a radar-view of η^2 from the first sampling. It's obvious here that SR and $pcdV$ are mostly impacted by the nativity of the speaker, and that $VarcoV$, n and $rPVI\_dV$ and σdV are partly affected by it.

5. Discussion

We suggest a computer model of French rhythm, trained on a large and varied speech data set on the basis of 16 acoustic duration parameters fully automatically detected from the signal.

This model showed a significant rhythmic difference between A2-level Japanese speakers of French and native speakers from the CEFC corpus. However, it showed a less significant difference between natives and non-natives from the CEFC. Many reasons may explain this smaller difference, such as heterogeneity of French proficiency levels of non-natives in the CEFC corpus as well as heterogeneity of their mother tongues. Rhythmic scores of proximity to the model were correlated to French global proficiency level for the Japanese speakers, yet we still need to include more heterogeneous proficiency levels in our non-native test set to better test our model. Also, correlation remains less strong since it is a global proficiency level and not especially a pronunciation evaluation. It would be interesting now to compute the correlation between rhythmic scores and results of a perceptual test of foreign accent degree on the same speakers.

What we learned with the acoustic parameters efficiency analysis, is that after speech rate, vowels and voiced consonants' duration seems to be one phenomenon that clearly differs between natives and Japanese speakers of French. Speech needs to be speeded up, with fewer silences and more variation in voicing duration. On the other hand, intersyllabic duration seems not to be such an important factor for differentiating between NS and JpNNS production. This might be linked to the fact that both French and Japanese are syllable(/mora)-timed languages. The next step of this study will be to compute rhythmic proximity of other non-native speakers, from various French proficiency levels and various mother tongues, then we might see to what extent rhythm is correlated to language proficiency and mother tongues, and what are the most efficient rhythmic parameters to characterise foreign accent depending on learners' mother tongues.

6. Acknowledgements

This work was carried out during an internship in Laboratoire d'Informatique de Grenoble (LIG) of Université Grenoble Alpes in 2019, in the Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole (GETALP).

We thank the EMERGENCE project of LIG for making this work possible, our colleagues of GETALP for their support, as well as Romain Jourdan-Ōtsuka from Kyōto University of Foreign Languages for providing the Japanese students corpus.

7. Bibliographical References

Alazard, C. (2013). *Rôle de la prosodie dans la fluence en lecture oralisée chez des apprenants de Français Langue Étrangère*. Ph.D. thesis, Université Toulouse 2. Thèse de doctorat dirigée par Michel Billières et Corine Astesano.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, 66 1-2:46–63.

Astésano, C., (2016). *Prosodic characteristics of Reference French*, chapter 5. Oxford University Press, oxford scholarship edition.

Benzitoun, C., Debaisieux, J.-M., and Deulofeu, H.-J. (2016). Le projet orfÉo : un corpus d'études pour le français contemporain. *Corpus*, 15:91–114.

Bhat, S., Hasegawa-Johnson, M., and Sproat, R. (2010). Automatic fluency assessment by signal-level measurement of spontaneous speech. *Second Language Studies: Acquisition, Learning, Education and Technology*, 01.

Boersma, P. and Weenink, D. (2019). Praat: doing phonetics by computer [computer program]. Version 6.0.37, téléchargée en mars 2019 depuis `http://www.praat.org/`.

Coulange, S. (2019). Proximité rythmique entre apprenants et natifs du français – évaluation d'une métrique basée sur le cefc. Master's thesis, Université Grenoble Alpes, 08. Mémoire de master dirigé par Solange Rossato en master de Sciences du langage parcours industries de la langue.

De Jong, N. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390.

De Meo, A., Pettorino, M., and Vitale, M. (2012). Comunicare in una lingua seconda. il ruolo dell'intonazione nella percezione dell'interlingua di

apprendenti cinesi di italiano. In *La voce nelle applicazioni. Proceedings of the 7th Congress of Italian Association of Speech Sciences AISV*, pages 117–129.

Dellwo, V. and Fourcin, A. (2013). Rhythmic characteristics of voice between and within languages. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 59:87–107.

Dellwo, V., Leeman, A., and Kolly, M.-J. (2015). Rhytmic variability between speakers: Articulatory, prosodic and linguistic factors. *The Journal of the Acoustical Society of America*, 137(3).

Di Cristo, A. and Hirst, D.). L'accentuation non emphatique en français : stratégies et paramètres. In *Polyphonie pour Ivan Fónagy*, pages 71–101. L'harmattan edition.

Dister, A., Francard, M., Hambye, P., and Simon, A.-C. (2009). Du corpus à la banque de données. du son, des textes et des métadonnées. l'évolution de banque de données textuelles orales valibel (1989-2009). *Cahiers de Linguistique*, 33/2:113–129.

Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., and Precoda, K. (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication*, 69(C):31–45, 05.

Flege, J. (1988). Factors affecting degree of perceived foreign accent in english sentences. *The Journal of the Acoustical Society of America*, 84(1):70–79, July.

Fontan, L., Le Coz, M., and Detey, S. (2018). Automatically measuring l2 speech fluency without the need of asr: A proof-of-concept study with japanese learners of french. In *Interspeech 2018*, pages 2544–2548, 09.

Fourcin, A. and Dellwo, V. (2013). Rhythmic classification of languages based on voice timing. *Tranel Review*, pages 87–107, 07.

Gibbon, D. and Gut, U. (2001). Measuring speech rhythm. In *EUROSPEECH 2001*, pages 95–98, 01.

Grabe, E. and Low, E. L., (2002). *Durational variability in speech and the rhythm class hypothesis*, volume Vol. 7, pages 515–546. Mouton de Gruyter, 01.

Pellegrino, E. (2012). The perception of foreign accented speech. segmental and suprasegmental features affecting degree of foreign accent in italian l2. *Mello H. et al. (Eds.) Proceeding of the 8 GSCP Conference*, pages 261–267.

Pettorino, M., Maffia, M., Pellegrino, E., Vitale, M., and De Meo, A. (2013). *VtoV: a perceptual cue for rhythm identification.* University of Leuven (KU Leuven).

Piske, T., MacKay, I., and Flege, J. (2001). Factors affecting degree of foreign accent in an l2: a review. *Journal of Phonetics*, 29(2):191 – 215.

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73:265–292, 12.

Rossato, S., Zhang, D., Ajili, M., and Bonastre, J.-F. (2018). Suivre le rythme de tes paroles. In *Proc. XXXIIe Journées d'Études sur la Parole*, pages 37–45.

White, L. and Mattys, S. (2007a). Calibrating rhythm: First language and second language studies. *J. Phonetics*, 35:501–522.

White, L. and Mattys, S. (2007b). Rhythmic typology and variation in first and second languages. In *Segmental and prosodic issues in Romance phonology*, pages 237–257. John Benjamins Publishing Company.