# Evaluating prosodic similarity as a means towards L2 teacher's prosodic control training

*Olivier Nocaudie, Corine Astésano*

U.R.I Octogone-Lordat (E.A. 4156), Université de Toulouse, UTM, Toulouse, France

nocaudie@univ-tlse2.fr, corine.astesano@univ-tlse2.fr

## Abstract

Studies on professional impersonators and naïve speakers has underlined that speech imitation proficiency varies across speakers. Imitation in speech supposes that a speaker succeeds in reproducing specific features of the perceived speech. Because of the inherent variability of human speech behaviors, the question lies open whether different speakers can accurately imitate phonetic features, and more specifically prosodic patterns. This exploratory study proposes to test *f*0 contours' imitation of 4 sentences originally pronounced by a female speaker, by 4 naïve listeners undertaking 3 different tasks: mere repetition, imitation and exaggeration of the perceived sentences., Two tests were performed: imitated sentences and models were time-warped and objective comparisons were performed using two (dis)similarity measures reported in the literature; a panel of 15 listeners evaluated perceptually the same set of sentences during an AX similarity judgment task. Similarity scores were used to build multiple rankings in order to observe the correlation between the two tests' rankings and to evaluate prosodic imitation proficiency across speakers/listeners. This research has implication for L2 phonetic correction using the Verbo-Tonal Method, which requires excellent prosodic awareness and control by the teacher in the production of lexicalized and delexicalized sentences.

**Index Terms**: speech imitation, prosodic patterns, time-warping methods, perception of prosodic similarity.

## 1. Introduction

Studies on speech imitation report different types of human imitation behavior such as convergence (mutual adaptation during the course of the interaction) [1], voice disguise (attempt of impersonating someone else's voice) [2], [3], or mere imitation (simple mimicry [4], shadowing [5], [6]). On the one hand, these behaviors may be defined as different in so far as they depend on factors like contexts of production [7] or imitator's intention [8]. On the other hand, they share a major common trait, namely their qualification as imitative speech behavior: the speaker's production must sound similar to its model, whatever imitation characteristics are used. Thus, speech imitation's studies aim to elicit and observe behavioral shifts in the way speakers talk, may it be at a lexical or a phonetical level. For the latter, uncovering what feature in the signal is being imitated and how it is being assessed often remains a methodological puzzle. Indeed, it is delicate to choose what acoustical features to measure and to link to the results of perceptual tests [9].

Professional impersonators tend to use global adjustment to the voice's target specificities and are also able to imitate instant variations (synchrony strategies) like intonation contours or duration of pauses. Naïve speakers, however, seem to be limited to convergence strategies (global adjustment to the voice) [2] [3]. This observation raises some interconnected questions related to synchrony strategies: (1) To what extent can a naïve speaker reproduce a perceived prosodic pattern (instant variations); (2) How can we assess their success or failure in doing so; (3) Is it possible to train a speaker to reproduce intonation, and more generally any prosodic feature, more accurately?

Questions (1) & (3) have a specific relevance in the domain of teaching pronunciation to L2 speakers, more particularly in the framework of the Verbo-Tonal Method (hereafter VTM). VTM postulates that errors of pronunciation in L2 are due to a L1 bias in the perception of the L2. To neutralize the effect of this bias, VTM proposes to exercise the speaker's ear using a wide array of correction processes where prosody has a crucial role. A teacher using VTM must have specific prosodic awareness and control, more particularly when (s)he is required to delexicalize (or logatomize) a sentence in order to facilitate the perception of the rhythmic and intonational features of the target language, by drawing the learners' attention on these prosodic features.

Per se, live phonetic correction performance represents a typical imitative interaction. Indeed, during VTM interaction, both the teacher and the trainee have to imitate or reiterate some speech features. The trainee (L2 learner) has to repeat the teacher's linguistic model, which leads one to question the link between speech perception & (re)production in bilingual learners. The teacher has to produce coherent phonological and prosodic patterns consistently, which raises the question of production control, more specifically at a prosodic level.

If questions (1) and (3) may apply to both the teacher and the trainee, the present study focuses on the teacher's aptitude to consistently reproduce prosodic patterns. Indeed, before addressing learners' ability to imitate/(re)produce linguistic features, one has to make sure that the *imitee* (L1 teacher) is actually able to consistently reproduce (hence, imitate) his/her own speech. As mentioned before, phonetic correction in the VTM framework implies repetition of prosodic features in a consistent way; it also implies that the teacher is able to emphasize some prosodic realizations to facilitate learners' perception of the target features. We therefore propose to first test L1 speakers' ability to control their imitation of prosodic features. In doing so, we address question (2), *i.e.* assessing for speakers' success or failure at imitating prosodic features. Ultimately, the methods used to assess prosodic (dis)similarity is intended to evaluate teachers' prosodic control and be used as a tool for their training.

 doi: 10.21437/SpeechProsody.2016-6

Few studies have tackled the issue of speech imitation in French, and more specifically on prosodic cues' imitation (see however [10] for Initial Accent reproduction). The present study is following up on our previous preliminary study describing speakers' ability to imitate prosodic features of controlled sentences on an 'imitation scale' going from a simple repetition to an exaggerated mimicry [11].

## 2. Linguistic material: An imitation corpus

The corpus originally consists of syntactically ambiguous sentences that can be disambiguated via prosodic cues. Syntactic ambiguity derives from the manipulation of the adjective scope on two coordinated nouns, as in "les gants et les bas lisses" (*the smooth gloves and stockings*), where the adjective (A) "lisses" either qualifies the second noun "bas" only ([les gants][et les bas lisses]; Low Adjective attachment hereafter *Low*), or either the two nouns "gants et bas" ([les gants et les bas][lisses]; High Adjective attachment, hereafter *High*). Sentences vary in terms of Noun and Adjective lengths, from one to four syllables. Manipulating syntactic ambiguity and constituents' lengths allows us to uncover the prosodic cues (prominences, boundary tones, pauses …) used for syntactic linearization of spoken utterances. For more details on this corpus, see [12].

A subset of 16 sentences spoken by a female speaker was selected for our imitation tasks. These sentences consisted of two Noun lengths (tri- and quadri-syllable nouns) combined with one- to four-syllable lengths of Adjective, in the two syntactic readings conditions (*Low* and *High*). 8 native listeners/imitators of French were instructed to speak out sentences in three different tasks performed in separate blocks: a) a mere repetition *(Rep*); b) an imitation (*Imi*); and c) an exaggerated imitation (*Exa*) of the speaker's sentences. 2 speakers were discarded for voice quality problems or experiment-induced stress. In each block, listeners/imitators repeated each sentence 3 times, in a random order, giving rise to a total of 864 sentences (16 sentences * 2 syntactic conditions * 3 repetitions * 3 tasks * 6 speakers). In order to evaluate the implicit ability to imitate speech, the attention of the speakers was not drawn to imitation in task a); they were instructed to just "*say the sentence while preserving the intended structure*". In tasks b) and c), they were explicitly asked to imitate and to exaggerate the sentences. Their attention was however not drawn to prosodic features.

In the present exploratory study comparing objective and subjective data, we chose to select a subset of 4 sentences from this corpus according to two criteria chosen to evaluate the robustness of the algorithm used to test prosodic similarity: 1) the sentences were all taken from the *Low* attachment syntactic condition because syntactic disambiguation is marked by a silent pause between the first and the second noun. The presence of acoustic silence is of particular interest to test for robustness insofar as the algorithm is overly biased towards silence alignment when evaluating prosodic similarity; 2) the sentences were chosen to illustrate two different phrase lengths. We also chose to run the present tests on 4 listeners/imitators only (Sp1, Sp3, Sp5 and Sp7), who were paired to imitate the following sentences:

- Sp1 (female) & Sp5 (female)
    - Les bagatelles et les balivernes sottes
    - Les bonimenteurs et les baratineurs fades
- Sp3 (male) & Sp7 (female)
    - Les bagatelles et les balivernes saugrenues
    - Les bonimenteurs et les baratineurs fabuleux

Altogether, our results will be computed on 18 sentences by subject, yielding a total number of 72 sentences ([2 sentences * 3 repetitions * 3 tasks] * 4 subjects).

## 3. Method: Objective measurements & perceptual evaluations of prosodic imitation

Section 3 describes our methodology for evaluating imitated *f0* contours (dis)similarity with our speaker's model. It also presents the perceptual evaluation task that was undertaken for comparison with the objective measurements.

One problem raised by the assessment of imitation in speech lies in the absence of congruence between perceptual judgments of imitation and the multitude of acoustic features either converging with or diverging from the model [7], [9].

Pitch, and its physical correlate *f0* is reported to be the main feature targeted by imitators [3]. It is also the primary cue used for corrective feedback during VTM correction. Our method will thus focus on the measurement of the physical distance between pairs of *f0* contours on the one hand, and on the perceptive evaluation of their resemblance on the other hand.

### 3.1. Dynamic Time-Warping (DTW) & (dis)similarity measures

Assessing imitation of *f0* contours objectively amounts to find if there is a physical distance between these contours, *i.e.* to answer the question of the shapes' matching of the contours.

Shape matching however supposes tonal normalization and temporal alignment of *f0* peaks and valleys (DTW). The distance between two *f0* contours was computed through two measures similar to the method proposed by Hermes [13] where *w(t)* is the temporal course of the weighting factor (*i.e.* the sum of the reference signal's subharmonic sumspectrum), *W* its time integral from *0* to *T* (*T* being the duration of the utterance), $f_1$ and $f_2$ the tested pitch contours of sentence pairs.

We however chose to use a different normalization procedure than that of Hermes, and divided each *f0* values by the maximum *f0* of the utterance ($f_1=p_1/p_1max$). This normalization procedure allows for comparing male and female listeners/imitators by bringing *f0* variations on a comparable scale from zero to one, relative to speakers' mean *f0*. This will later help peaks' and valleys' comparisons using the DTW algorithm's comparison. The sampling rate was one *f0* values per millisecond [14] extracted with Praat [15].

After normalization, the root mean square difference ($L_2$) between two contours was computed as follows:

$$L_2 = \left\{\frac{1}{W}\int_0^T w(t)\left|f_1(t) - f_2(t)\right|^2 dt\right\}^{1/2} \quad (1)$$

A correlation coefficient (*r*) between the two contours $f_1$ and $f_2$ was then computed as follows:

$$r = \frac{\frac{1}{W}\int_0^T w(t) f_1(t) f_2(t) dt}{\sqrt{\left\{\frac{1}{W}\int_0^T w(t)\left|f_1(t)\right|^2 dt \frac{1}{W}\int_0^T w(t)|f_2(t)|^2 dt\right\}}} \quad (2)$$

Hermes [13] however reports that *r* needs to be transformed in Fischer's *Z* (hereafter *Zr*) to allow for correlation's comparison:

$$Z_{f_1 f_2} = \frac{1}{2} ln \frac{1+r_{f_1 f_2}}{1-r_{f_1 f_2}} \quad (3)$$

$L_2$ measures rapid changes in the f0 contour while *Zr* is a holistic measure of contour shapes.

Before computing $L_2$ and *Zr* for each pair of *f0* contours, Dynamic Time Warping was performed on the tested contours to force the alignment between the model and its reproduction (non-linear *f0* interpolation). It has been reported that such an alignment would overall improve the correlation, especially when the contours are functionally similar, *i.e.* when they share the same accentual pattern [14].

Finally, each sentence was ranked relatively to the others, depending on their $L_2$ and *Zr* scores:

- $L_2$ is a **dissimilarity measure** (the higher the $L_2$, the higher the dissimilarity). The sentence with the lowest $L_2$, was ranked 1 while the one with the highest $L_2$ was ranked 72.
- *Zr* is a **similarity measure** (the higher the *Z*, the higher the similarity). The sentence with the highest *Z* was ranked 1, the second highest was ranked as 2, and so on.

Two objective rankings were thus obtained, which will be compared to the ranking derived from the perceptual evaluation's results (see 4.3.2).

### 3.2. AX similarity judgment test

As argued by [9], imitation in speech should be assessed both objectively and subjectively, *i.e.* physically and perceptively. To this end, we complemented the objective measurements described above with an AX similarity judgment task, which allows for an absolute rating of each reiterated sentence (X) compared to the model (A). 15 naive listeners participated to the AX judgment task. All were French native speakers (age 25-32) and did not report any hearing or speech disorder.

Listeners were instructed to rate the resemblance of X with A in terms of the 'musical' features of speech (rhythm, tonal variations). The task was run on a computer using the Lancelot software (HTML environment of PERCEVAL [16]). Sentences were randomized by the software and auditorily presented using high quality headphones. Listeners could hear each pair of sentences up to five times before giving their rating on a scale from 1 (less similar) to 5 (perfect match) by clicking on the corresponding button with the computer mouse.

Results were computed by calculating the mean score of each X sentence. In case of ties, we attributed to groups of tied sentences a rank equal to the mean of their consecutive original ranking.

## 4. Results

#### *4.1.1. Distribution of objective and perceptual rankings*

We first describe the results comparing the rankings obtained from objective scores ($L_2$ and *Zr*) and the perceptual scores (AX) across speakers. Figure 1 shows the distribution of the 3 different scores, which will give rise to the calculation of correlation coefficients. Box plots show the global mean ranking (dots) and the interquartile distribution of the ranks by subject on 18 sentences (note here that the 3 different imitation tasks are merged for now). Whiskers indicate minimum and maximum ranking values. Lowest mean rankings indicate better judgment in *f0* contour comparisons.

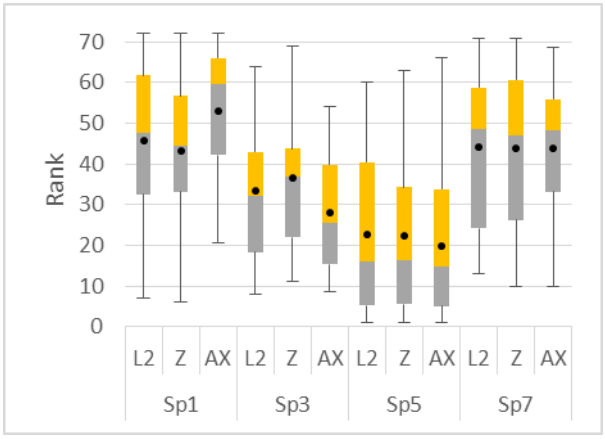


Figure 1: *Distribution of* $L_2$, Zr *and AX ranks per subject (Sp1, Sp3, SP5, Sp7), classified from 1 to 72 sentences (y axis). Dots represent the mean ranks over the 18 sentences produced by each speaker.*

Given this distribution of ranks across the subjects, it seems that Sp5 can be classified as the most proficient subject (Mean rank $L_2$ = 22,76, *Zr* = 22,28 and AX = 19,88) followed clearly by Sp3 ($L_2$ = 33,33, *Zr* = 36,56, AX = 27,92). Objective rankings of Sp7 ($L_2$ = 44,17, *Zr* = 43,83) and Sp1 ($L_2$ = 45,83, *Zr* = 43,33) are close to each other, but their perceptual rankings (respectively (AX(Sp7) = 43,82 ; AX(Sp1) = 53,06) may reflect the dispersion of their ranks in the inferior quartile: Sp1's best rank is greater than Sp7's, but it may act as an outlier for the computation of their mean score. Overall, Sp7 obtained a greater amount of good ranks than Sp1 during every evaluation process task, as shown in the box plot.

According to Hermès [13], $L_2$ measures the perceptual distance between two contours, where quadratically more weight is given to larger distance. *Zr* expresses the distance between the contours' shape, *i.e.* to what extent can a pitch contour be obtained from another by performing a linear transformation. Given their different nature, it is of interest to correlate them both with perceptual evaluation results in order to later determine a threshold on $L_2$ and/or *Zr* beyond which we could estimate fairly accurately the results of perceptual judgement of prosodic similarity. Figure 2 show the correlation between $L_2$ and *Zr* scores for the 72 sentences. The points in the lower right corner represent sentences estimated as highly similar with the model.

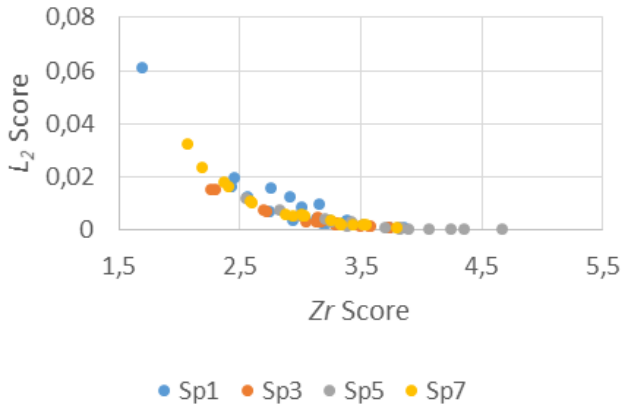


Figure 2: $L_2$ *(RMS difference) and* Zr *scores (Fisher transform of r) of the 72 sentences.*

#### 4.1.2. *Correlation between Zr, $L_2$ and AX rankings*

The correlation was computed using *Real Statistics* for Excel [17]. A test of correlation between objective and subjective measures based on *ordered* data (ranks) is possible when using the $r_s$ of Spearman, which allows for the comparison between different rankings. Pairwise two-tailed tests show fairly good correlation of *Zr* ranking with AX ranking ($r_s$ = .554, $p$ < .0001, $t$ (71) = 5.562), while the correlation between $L_2$ and AX rankings were slightly stronger ($r_s$ = .589, $p$ < .0001, $t$ (71) = 6.092). Both correlation values are indeed above $r_s$'s critical value for $N$ = 72 ($r_s$-*crit* = .382; *t-crit* = 3.43). The linear relationship between objective and perceptive rankings thus seems pretty robust.

#### 4.1.3. *Imitation tasks and performance*

Results given by the algorithm underline the difference of imitation proficiency across speakers/listeners. Figure 3 illustrates proficiency differences between the two paired speakers which respectively are the less (Sp1) and the most (Sp5) proficient in the tasks, as rated both by the algorithm and the panel of listeners. We predicted that the more conscious imitations (tasks *IMI* and *EXA*) would be produced as most prosodically accurate. However, both objective and perceptual results indicate great imitation performance variation across speakers. Whereas Sp5 seemingly shows a better control with increasing performance throughout the three tasks, some of Sp1's *REP* sentences exhibit better rating than other sentences produced during the *EXA* task. Note that bigger dots in the lower right corner indicate perceptively better rated imitations

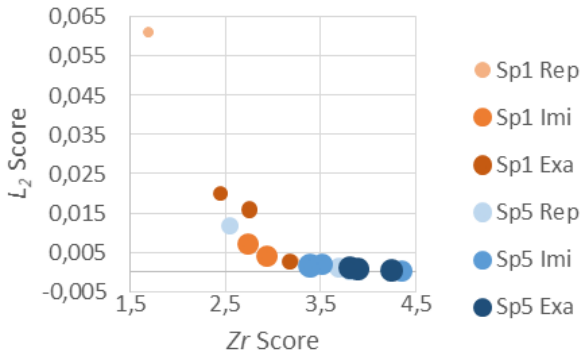


Figure 3: *Illustration of speakers' proficiency in performing the 3 imitation tasks for 9 repetitions of sentence 'Les bagatelles et les balivernes sottes'.* L2 *and* Zr *scores are on the y and x axes; AX judgments' mean scores are represented by bullets' size.*

## 5. Discussion

This preliminary study was intended to methodologically test the validity of comparing objective and perceptual evaluations of prosodic similarity in imitation. Our ultimate goal is to find a sufficiently robust algorithm method, which could be implemented as an automatic tool evaluating teachers' proficiency at imitating prosodic contours.

Originally, the DTW was used by [18] as a measure of convergence in speech, expressed as *d*, the output of DTW giving the cost of alignment between two contours. With our aim in mind, we chose here to use the DTW as an interpolation method (as proposed by [14]) to compute two measures of prosodic similarity (initially reported by [13]): the first measure ($L_2$) models rapid perceptual processes, while the second measure (*Zr*) models holistic perceptual processes related to contours' shapes.

Both measures correlated well with the perceptual test performed on 15 listeners for this exploratory study: bad and good imitations were consistently spotted by the algorithm too. The difference of correlation between $L_2$ and *Zr* may reflect the nature of both measures, as discussed earlier. That being said, these results encourage us to elaborate on the automatic investigation of imitation but the question lies open whether both $L_2$ and *Zr* measures are to be kept in a near future to continue our experiments. In other words, studies on a larger database are necessary to set a threshold on $L_2$ and/or *Zr* beyond which the results of perceptual judgement are accurately enough estimated and to evaluate potential discrepancies between these two factors of similarity.

As our first results were encouraging, it is planned to expand this approach in multiple directions:

- More sentences from the imitation corpus will be rated, both objectively and subjectively.
- A new corpus consisting of sentences and their delexicalized reproductions by human speakers will be constituted, in order to further test these types of measures.

Besides, it may be of interest to test a method of shape matching involving a different transformation than DTW, which requires thousands of *f0* values, and quite a long computing cost. Among the methods of shape matching reviewed by [19], the cumulative angle function could be applied to *f0* contours, stylized with the help of much fewer sampling points. It could lead to refine prosodic patterns analysis (slopes and timing), which might be a satisfactory substitution to the *Zr* measure. Ultimately, it is intended to limit the use of perceptual tests in the assessment of prosodic imitation in speech, by selecting the factors of objective similarity best correlating with extensive perceptual results.

The tasks performed to gather the corpus intended to underline the capacity of naïve speakers to imitate, in the same way a prosodically unaware teacher could do when trying to correct phonetics of L2 learners. For some speaker (as Sp1) the algorithm may help diagnose if they exhibit or not prosodic awareness and control, and to some extent, talent. As underlined by [20], talent, as an individual factor is complex to assess. This type of objective approach could be used to detect that part of the talent of individuals resorting to prosodic ability.

Finally, our perspective will be to focus on specific training of VTM, more precisely, on the correctness of prosodic cues reproduction. Ideally, our research should lead to build a user-interface allowing teachers to train specific VTM processes, in this case, delexicalization used to help focus on syllabification and rhythm.

## 6. Acknowledgements

This study is supported by the Agence Nationale de la Recherche grant ANR-12-BSH2-0001 (PI: Corine Astésano)

We would like to thank Albert Rilliard, LIMSI, CNRS, France, for his advices on the topic of prosodic similarity comparison; and Benjamin Boulbène & Julien Dupouy, France for their involvement with implementing the algorithm.

## 7. References


[1] J. S. Pardo, « On phonetic convergence during conversational interaction. », *J. Acoust. Soc. Am.*, vol. 119, nº 4, p. 2382‑93, 2006.

[2] E. Zetterholm, « A comparative survey of phonetic features of two impersonators », in *Fonetik*, 2002, vol. 44, p. 129–132.

[3] J. Revis, C. De Looze, et A. Giovanni, « Vocal Flexibility and Prosodic Strategies in a Professional Impersonator », *J. Voice*, vol. 27, nº 4, p. 524.e23‑524.e31, juill. 2013.

[4] H. Mixdorff, J. Cole, et S. Shattuck-Hufnagel, « Prosodic Similarity–Evidence from an Imitation Study », in *Speech Prosody 2012*, 2012.

[5] S. D. Goldinger, « Echoes of echoes? An episodic theory of lexical access. », *Psychol. Rev.*, vol. 105, nº 2, p. 251‑279, 1998.

[6] S. Dufour et N. Nguyen, « How much imitation is there in a shadowing task? », *Front. Psychol.*, vol. 4, 2013.

[7] N. Lewandowski, « Talent in non-native phonetic convergence », Universität Stuttgart, Stuttgart, 2012.

[8] M. Donald, *Origins of the Modern Mind - Three Stages in the Evolution of Culture & Cognition*, Reprint. Cambridge, Mass.: Harvard University Press, 1993.

[9] J. Pardo, « Reconciling diverse findings in studies of phonetic convergence », in *Proceedings of Meetings on Acoustics*, 2013, vol. 19, p. 060140.

[10] A. Michelas et N. Nguyen, « Uncovering the Effect of Imitation on Tonal Patterns of French Accentual Phrases. », in *INTERSPEECH*, 2011, p. 973–976.

[11] O. Nocaudie et C. Astésano, « Prosodic structuring imitation in French L1 context-A first step towards correcting phonetic-prosodic features in L2 French », in *Proceedings of ISICS*, Aix-en-Provence, 2012.

[12] C. Astésano, E. G. Bard, et A. Turk, « Structural influences on Initial Accent placement in French. », *Lang. Speech*, vol. 50, nº 3, p. 423‑446, 2007.

[13] D. J. Hermes, « Measuring the Perceptual Similarity of Pitch Contours », *J. Speech Lang. Hear. Res.*, vol. 41, p. 73‑82, 1998.

[14] A. Rilliard, A. Allauzen, et P. B. de Mareüil, « Using Dynamic Time Warping to Compute Prosodic Similarity Measures. », in *INTERSPEECH*, 2011, p. 2021–2024.

[15] P. Boersma, « Praat, a system for doing phonetics by computer. », *Glot Int.*, vol. 5:9/10, p. 341‑345, 2001.

[16] C. André, A. Ghio, C. Cavé, et B. Teston, « PERCEVAL: a Computer-Driven System for Experimentation on Auditory and Visual Perception », in *Proceedings of XVth ICPhS*, Barcelone, Espagne, 2003, p. 1421‑1424.

[17] C. Zaiontz, *Real Statistics Using Excel*. 2015.

[18] M. Kim, « Phonetic accommodation after auditory exposure to native and nonnative speech », NORTHWESTERN UNIVERSITY, 2012.

[19] R. C. Veltkamp, « Shape matching: similarity measures and algorithms », 2001, p. 188‑197.

[20] M. Jilka, H. Baumotte, N. Lewandowski, S. Reiterer, et G. Rota, « Introducing a comprehensive approach to assessing pronunciation talent », *Proc. 16th ICPhS Saarbr.*, p. 1737–1740, 2007.